Extracting and Modeling the Geography of Text Documents

Resources and Applications

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Elise Anne Acheson
von Grossaffoltern BE

Promotionskommission
Prof. Dr. Ross Purves (Vorsitz)
Prof. Dr. Robert Weibel
Prof. Dr. Christopher Jones

Zürich, 2019

Abstract

Spatial language is an important part of everyday communication, and placenames, like Zürich and Switzerland, are a major way to refer to geographical locations. In written text documents, these placenames and their associated spatial language convey spatial information to the reader, setting the geographical context of news, events, blogs, and even scientific research. For the intended audience of a text, shared experience and spatio-temporal context makes understanding these references to places effortless. However, for computer systems, automatically extracting and representing this geographical content is a challenging process which must deal with placename ambiguity and vagueness in spatial language. Indeed, creative strategies are required to turn text about locations into information that is explicitly spatial, such as points or regions that can be visualized on a map.

It is this **text-to-space** process which is the subject of this thesis, focusing both on *resources* used to link placenames and geographical representations, and on *applications* of text-to-space pipelines for particular use cases. Placenames form the 'glue' to create geographical representations for text documents, because they can be identified in text using tools like Named Entity Recognition (NER), and linked to spatial representations (geometries) which are catalogued in placename resources known as gazetteers. In their simplest form, gazetteers provide a name, feature type, and geometry for a set of named places. However, each gazetteer differs in terms of which places are catalogued and how, with no global authoritative gazetteer providing a definitive list.

In the first part of this thesis, we look at this gazetteer heterogeneity, first by analyzing and comparing the spatial *coverage* of two global gazetteers using aggregated record counts, and then by aligning or *matching* individual records from two gazetteers when they represent the same real-world entity. In our analysis of gazetteer coverage, we find wide discrepancies in coverage between the two global gazetteers we compare, with the main driver of variation being the country unit, and particularly unbalanced and idiosyncratic coverage for common natural feature types. In our work on gazetteer matching, we present a detailed machine learning pipeline to match near-duplicate natural feature records across two gazetteers using a random forest classifier, and compare these machine learning results to rule-based matching. We find that machine learning outperforms our best rules by about

6%, adapts better to different feature types, and performs increasingly well as we use more training examples. Our matching pipeline could be applied to integrate placename resources for particular text-to-space applications.

In the second part of this thesis, we build and apply text-to-space pipelines for two case studies, focusing on understudied text types and on real-world applications. The first case study builds a semi-automatic pipeline to create areal footprints for a set of hiking blogs, as part of a wider study on how people describe landscapes in Switzerland. Our pipeline converts manually annotated toponyms to a set of points found in a gazetteer, then removes spatial outliers using filtering and clustering, before generating polygons (convex hulls) around the remaining points. Our second case study builds a fully-automatic pipeline to extract and represent relevant geographic information from scientific articles in two different domains: the biomedical domain, where relevant locations are typically patient treatment locations, and the ecological domain, where relevant locations are predominantly field study sites. We report results in terms of precision and recall and produce a global map for each corpus. Our detailed error analysis suggests that performance improvements will likely result from improving various individual pipeline components. The outputs of our pipeline could be used in a meta-analysis or to geographically search or filter articles.

Considering the sum of our work, we offer a list of recommendations for building a text-to-space pipeline for a new application, based on properties of the text corpus and on task requirements. Future work could systematically evaluate these recommendations, as well as relate changes in the tools and resources used in a pipeline to performance on a task. Other directions for future work include varying the types of texts processed, expanding the size of the corpora and making more efficient pipelines, and experimenting with how best to customize gazetteer resources for a particular task. In conclusion, this thesis contributes theoretical and applied knowledge about geographically modeling text documents through text-to-space pipelines.

Contents

List of Abbreviations

API . . . . . Application Programming Interface.

DBSCAN . . . Density-Based Spatial Clustering of Applications with Noise (a clustering algorithm).

GIR . . . . . Geographic Information Retrieval.

NER . . . . . Named Entity Recognition.

NLP . . . . . Natural Language Processing.

POI . . . . . Point of Interest (a feature type encompassing restaurants, cafes, museums, and so on).

POS . . . . . Part-of-Speech (such as noun, verb, or adjective).

TGN . . . . . The Getty Thesaurus of Geographic Names.

List of Figures

List of Tables

Part I

Synthesis

Though we experience space as continuous and three-dimensional, and time as continuous and inexorably flowing, there is nothing three-dimensional or flowing about expressions for space and time in language, which are staccato strings of sounds.

— Steven Pinker, The Stuff of Thought (Pinker, 2008)

1 Introduction

Contents

1.1 Motivation

Geographical locations are important in everyday communication, whether one is talking about where to meet, describing a recent trip, or discussing current events. When communicating about locations, names for places, known as placenames or toponyms, play a key role. Like names for people, placenames[1] are sequences of characters or sounds that single out one particular place to the reader or listener. Of course, like Peter, Paul, and Mary, placenames are seldom unique. However, in an everyday context, they are effortlessly understood by one's intended audience, based on shared experience and context. Perhaps the communication partners are in close proximity in space and time, or perhaps through interpersonal experiences the default sense of a placename was acquired and needs no further qualification when used. It may thus seem counter-intuitive that associating placenames with

[1] I will use the words 'placename' and 'toponym' interchangeably in this thesis, where 'toponym' is simply a more formal alternative to 'placename'.

particular geographical locations is quite a challenging process for a computer. Indeed, for computer systems, creative strategies are required to turn text about locations into information that is explicitly spatial, such as points or regions that can be visualized on a map.

It is this challenging *text-to-space* process which is the subject of this thesis, focusing both on ***resources*** used to link placenames and geographical representations, and on ***applications*** of text-to-space pipelines for particular use cases. The text-to-space process is about linking language - in our case, written language forming text documents - to geographical models. The key ingredient for this link are placenames, as these can be *identified* in many kinds of texts using existing tools, in particular Named Entity Recognition (NER) tools, actively developed by the Natural Language Processing (NLP) community. Identified placenames, potentially with accompanying spatial language, can then be individually *grounded* (that is, linked to geographical representations) using text-to-space resources commonly known as gazetteers, which in their simplest form are a list of placenames alongside a feature type (such as mountain or city) and a geometry (such as a point or a bounding box) (Hill, 2006). As a final step to link text documents to geographical models, document-level *geographical representations* can be computed using these gazetteer results, for example by augmenting or aggregating the returned geometries to better capture scale, and by filtering placenames deemed irrelevant for the document. These 3 steps represent a basic text-to-space pipeline, as illustrated in Fig. 1.1.

Figure 1.1: Overview of a typical 3-step text-to-space processing pipeline.

Many applications benefit from modeling the geographical content of text documents through such a text-to-space process. For example, geographic information retrieval systems need to assign spatial 'footprints' to documents in order for users to find information pertaining to specific regions of the world (Purves et al., 2007); news-aggregating websites may want to group news articles by geographic region or present them on a mapping interface (Teitler et al., 2008); travel bloggers may wish to automatically generate route maps for their excursion narratives (Moncla et al., 2014a); advertisers may want to match content to users based on geographical relevance; disaster relief efforts may want to pinpoint where to send support based on distress messages referring to specific places (Middleton et al., 2014). In all cases, enhancing a corpus of *implicitly* spatial text documents by adding a set of *explicitly* spatial representations (such as points or regions) opens up the corpus to spatial analyses, such as those which can be performed in a Geographic Information System (GIS) (Longley et al., 2005).

Gazetteer resources play a key role in this text-to-space process, most obviously as the way that explicitly spatial representations (geometries) are obtained for particular placenames in the grounding step. In addition, they are also used in the upstream placename identification step, to determine which strings in a text document constitute placenames, or more broadly, textual references to locations. Finally, they can help to disambiguate locations which share a name, for example by providing extra information about individual place records such as population, area, elevation, and so on. Gazetteers are increasingly being amalgamated to form global resources from regional or thematic resources, and are often accessed and queried via service-oriented architectures or Application Programming Interfaces (APIs). They thus form an object of study in themselves and constitute the first theme of this thesis: text-to-space *resources*.

Decisions about how to build and optimize a text-to-space pipeline are often closely linked to the particular use case or application. For example, the gazetteer resources should provide adequate coverage of the geographical regions and locations which are mentioned in the text corpus. Furthermore, other properties of the

text corpus, including how placename-rich documents are and how much they adhere to the rules of formal written language, may influence how certain tools and techniques perform, including for placename detection (e.g. NER tools) and grounding/disambiguation (e.g. Geocoding APIs). Hence, the second theme of this thesis is *applications*, explored through the building of customized text-to-space pipelines for two varied case studies.

In today's world of ubiquitous digitization, fast-breaking news, and international social networks, explicitly representing the implicit spatial context found in written texts is becoming increasingly important. Written texts can be disseminated to strangers halfway across the planet at the click of a button, effectively diluting the shared context between author and reader. Furthermore, content creators are now more than ever ordinary citizens rather than professional writers, and all face a potentially international audience that may not share their assumptions, know their conventions, or even speak their language. This new global reality intensifies the need to improve the automatic translation of spatial language into shareable, manipulable geographical representations, for diverse types of text documents and for a variety of applications.

1.2 Thesis overview

This thesis contains two major research themes: the first is the analysis and processing of text-to-space *resources* known as gazetteers, and the second is the building and customization of text-to-space processing pipelines for particular *applications*. Paper I (Acheson et al., 2017a) and II (Acheson et al., 2019) deal with the former theme, while paper III (Wartmann et al., 2018) and IV (Acheson and Purves, submitted) deal with the latter. The analysis of gazetteer resources is undertaken in paper I, where two global gazetteers are analyzed and compared in terms of their overall contents and a set of common feature types. In particular, their spatial coverage (placename density over geographical space) is quantified and compared at various scales of analysis, which tells us about their fitness-for-purpose for use in text-to-space pipelines. Continuing with the gazetteer resources theme,

paper II implements and compares rule-based and machine-learning-based methods to link individual gazetteer records across two different gazetteers when these records are deemed to be about the same real-world entity, a process referred to as gazetteer matching. Such a process could be carried out for a geographical area of interest prior to applying a text-to-space pipeline, in order to obtain better results, such as enriched placename information or increased recall in placename identification and grounding. Moving on to applications, papers III and IV build and apply text-to-space pipelines tailored to a particular use case. Paper III generates geographical representations, or 'footprints', from a corpus of informal landscape descriptions (hiking blogs) in order to enable the analysis of bottom-up landscape information gathered from different sources: in-person interviews, hiking blogs, and social media tags (Flickr photographs). Finally, paper IV builds a fully automatic text-to-space pipeline to extract and model relevant geographical information from scientific articles, such as study sites or patient treatment locations. This geographical information could then be used in a meta-analysis which considers the geographical context of the articles, or to retrieve articles by considering geographic relevance using techniques from Geographic Information Retrieval (GIR).

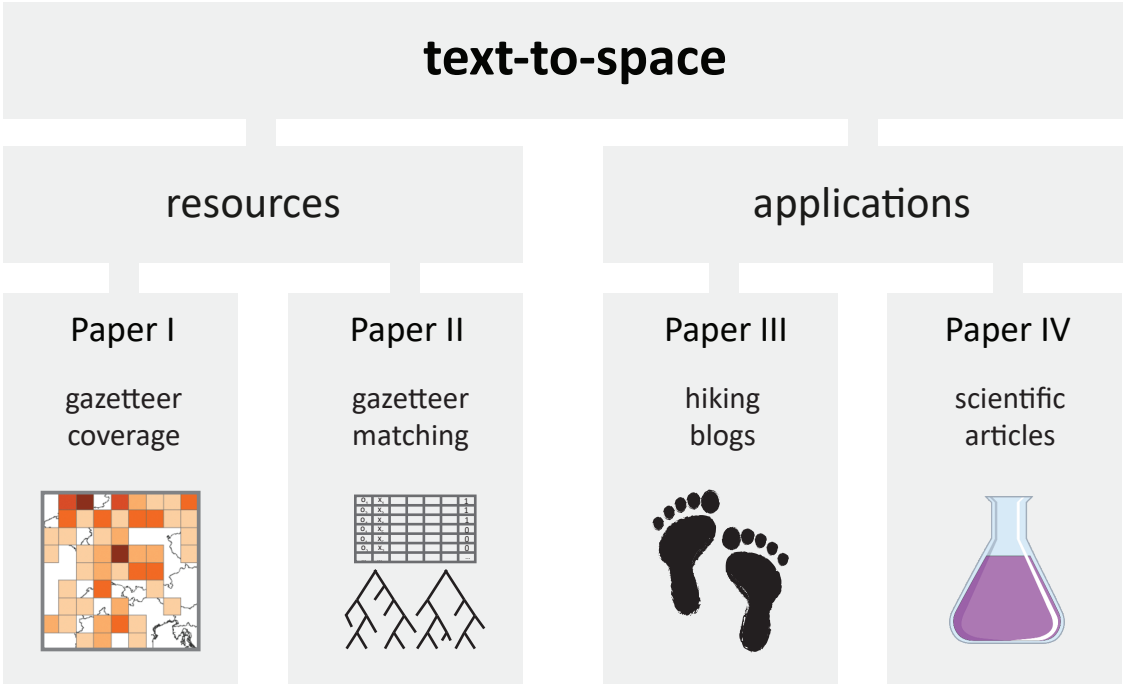


Figure 1.2: Overview of thesis themes and papers.

This thesis consists of two parts. Part I is the synthesis, including this introduction, followed by a concise review of the literature leading up to research gaps, presentation of the work undertaken in this thesis with one chapter for each of the two themes, and a discussion of what progress has been made and what remains. Part II consists of the 4 papers which have been written and published in the course of the PhD.

Alone on a train aimless in wonder
An outdated map crumpled in my pocket
I didn't care where I was going
They're all different names for the same place
The coast disappeared when the sea drowned the sun
I've no words to share with anyone
The boundaries of language I quietly cursed
And all the different names for the same thing

— Ben Gibbard, *Different Names for the Same Thing*,
Death Cab For Cutie

2 Background material

Contents

Connecting the discrete world of written words with the continuous domain of geographic space is a broad endeavour, with research contributions coming from a wide range of disciplines including Natural Language Processing, Geographic Information Science, Computer Science, and Cognitive Linguistics. In order to properly embed and support this thesis' contributions, it is necessary to provide, via relevant literature, a brief introduction to a few areas of inquiry, namely: spatial language, gazetteers, and text-to-space methods, including identifying placenames, grounding placenames, and geographically representing text documents. To make

this overview of literature more concrete, we first look ahead at some of the tasks which ultimately make use of document-level spatial models of text.

2.1 Tasks

Building geographic models of text documents via a text-to-space pipeline serves a variety of purposes. General tasks, alongside examples of concrete applications, include:

- **Information visualization**: Documents with any connection to places in the world can be displayed on a mapping interface for browsing and visualization. This can be done for various types of texts, including both for short factual[1] texts like news articles, for longer fictional narratives like books, and for historical texts of varying lengths. One example with news articles is the *NewsStand* application[2] which offers a map-based browsing interface for articles, clustering these based on both thematic and geographic content and presenting relevant articles to users based on their map extent (position and zoom level) (Teitler et al., 2008). On the literature front, books can be visualized on a map one at a time (Reuschel and Hurni, 2011), or the locations in a larger collection of literature can be mapped and interacted with, such as was done in the *Palimpsest* project for a corpus of literature set in Edinburgh, Scotland[3] (Alex et al., 2017). One of many works on historical texts looks at extracting and linking placenames near textual references to cholera in 19th century historical documents, allowing the corpus to be mapped and analyzed spatially (Murrieta-Flores et al., 2015).
- **Geographic Information Retrieval**: In order to return spatially relevant search results to users or answer queries of an explicit spatial nature (such as finding results pertaining to a particular region as depicted on a map), each indexed document (that is, each item to be potentially returned as a

[1] notwithstanding fake news
[2] interactive map at http://newsstand.umiacs.umd.edu/web/ (accessed in 06.2019)
[3] interactive map at https://litlong.org/ (accessed in 06.2019)

search result) should be assigned a relevant spatial representation (Larson, 1996; Purves et al., 2007; Purves and Jones, 2011).

- **Itinerary reconstruction**: Location-rich narrative text, such as travel blogs or hiking descriptions, can benefit from an overall spatial model to help with reconstructing routes, identifying and grounding references to locations, and inferring the location of uncatalogued and vague places (e.g. Moncla et al., 2014a; Budig and van Dijk, 2017).
- **User location determination**: The location of web users can be inferred using textual content - either on its own or as additional evidence in a model - to present them with geographically relevant content, such as news and advertisements. In this context, documents can be constructed by concatenating a user's written content such as micro-blog posts (Cheng et al., 2010; Mahmud et al., 2012) or search query logs (Gan et al., 2008).
- **Location information extraction**: Identifying textual references to locations in citizen-contributed content (including via social media or citizen engagement platforms) can help, for example, in a disaster relief context, to gather information about where help is needed or where damage has occurred (Middleton et al., 2014; Zhang and Gelernter, 2014), or in an urban planning and management context, to identify how citizens feel about different parts of their city or where attention to infrastructure is needed (Brando et al., 2016; Crooks et al., 2015). Jointly extracting places mentioned in text and information relating to those places can serve in a wide range of analyses, including to spatially analyze diseases from historical documents (Murrieta-Flores et al., 2015) and to characterize regions or landscapes based on how they are described in text (Derungs and Purves, 2013).

In addition to these examples of downstream tasks, document-level spatial models can also be used as a means to improve text-to-space processing itself, for example as an aid in placename disambiguation. Indeed, having an overall document geographic scope can help in reducing uncertainty about which is the correct referent for a placename, as evidence may lie in document-wide characteristics such as the

spatial distribution or granularity of textual locations and their potential referents (Smith and Crane, 2001; DeLozier et al., 2015; Hess et al., 2014).

As the above list of tasks shows, use cases for a text-to-space pipeline are quite varied and each may present its own trade-offs, such as comprehensiveness vs. display clutter in the case of visualization or precision vs. recall for geographic information retrieval. Hence, evaluation is important, not only for applications but also for the work on resources presented in this thesis. We thus now briefly look at the evaluation and similarity metrics used in this thesis.

2.2 Evaluation and similarity metrics

A range of evaluation and similarity metrics are used in this thesis, in diverse contexts including to evaluate how similar patterns of gazetteer coverage are, to evaluate how well placenames were extracted from a set of text documents, and to quantitatively assess how similar pairs of text documents are. The following list introduces key evaluation and similarity metrics which should facilitate comprehension of the work presented in Chapter 3 and Chapter 4:

- **Precision** is a heavily used metric in the contexts of both information retrieval and classification. The general idea behind precision is to determine how many positive instances found by some process (such as documents retrieved in an information retrieval task, or items classified as '1' in a binary classification task) are actually positive instances or 'true positives'. Hence, 'negatives', such as documents which are not retrieved and items classified as '0' by a classifier, are not considered by this metric. In a binary classification context, the formula to calculate precision is:

$$precision = \frac{TP}{TP + FP} \tag{2.1}$$

  where TP means 'true positives' (positives which were correctly classified as such) and FP means 'false positives' (when the process said 'positive' but the answer was 'negative'). In the context of finding correct matches in gazetteer

matching, precision can be stated as the number of positive matches correctly found over the total number of positive matches found:

$$precision = \frac{Positive\ matches\ correctly\ found}{Positive\ matches\ found} \tag{2.2}$$

- **Recall**, as opposed to precision, does concern itself with some 'negative' instances: those instances that were 'missed' by some process or falsely ended up in the 'negative' bin in a classification task. In the context of extracting relevant locations from text, a false negative would be a relevant location that was mentioned in the text but not extracted by the algorithm. Thus, here the denominator is not concerned with only what a process retrieves, as with precision, but instead is focused on what are the total instances *to be* retrieved - in other words, with the ground truth 'positive' instances. In the context of finding correct matches in gazetteer matching, recall can be stated as the number of positive matches correctly found over the total number of positive matches to be found (that is, positive matches in the ground truth):

$$recall = \frac{Positive\ matches\ correctly\ found}{Positive\ matches\ to\ be\ found} \tag{2.3}$$

- **Classification recall** specifies that, in a classification context, the instances considered in the recall calculation are only those actually seen by the classifier. In some multi-stage processes, such as gazetteer matching, some positive instances may not make it to the classification stage, but are still part of the overall ground truth. This subset of instances would factor into the overall 'recall' calculation for the entire process, but not into the 'classification recall' which would evaluate the classification in isolation. The formula to calculate classification recall in a binary context is:

$$recall = \frac{TP}{TP + FN} \tag{2.4}$$

where FN means 'false negatives'.

- **F1** is a very useful measure combining precision and recall through their harmonic mean, and thus summarizes the overall performance, since precision can typically be optimized at the expense of recall and vice-versa. The formula to calculate F1 is:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{2.5}$$

- **Accuracy** represents the percentage of correct classifications or predictions in general, considering both instances classified as positive and those classified as negative (as opposed to precision which considers only positive instances). Accuracy is useful when one can establish, for each prediction, whether it was 'correct' or 'incorrect', and then the overall performance can be summarized as the percentage of correct predictions over the total predictions:

$$accuracy = \frac{Correct\ predictions}{Total\ predictions} \tag{2.6}$$

  Accuracy is vulnerable to class imbalance, such as when most instances should be classified as 'negative' and thus high accuracy could be obtained by classifying all instances as negative. An example would be to classify each word in a text as 'location' (positive) or 'not a location' (negative). The vast majority of words in most texts will not be locations, and hence accuracy would be a poor way to judge this classification since one would easily obtain close to 100% accuracy just by classifying all words as 'not a location'. This is the main reason why precision, recall, and F1 are favored over accuracy in many contexts.

- **Kendall's tau**, also known as the Kendall rank correlation coefficient, is a measure of how similar two ordered lists are, taking values between 0 and 1, where 1 would be identically ranked lists. It is useful particularly for non-normally distributed data since it does not require a particular data distribution, only that the data can be ranked. One example of a ranked list is a list of countries ordered by how many gazetteer records each country contains. With two such lists, one can use Kendall's tau to compare how

similar the country counts in two gazetteers are (without considering the magnitude of the counts, but only their ordering).

- **Cosine similarity** is a measure of how similar two vectors are, which works by calculating the cosine of the angle between them. It can be used to compare how similar two text documents are, where each document is represented by a term-frequency vector, that is, a vector of length N where N is the number of words or 'terms' found across the two documents and each entry in the vector is the frequency of that word in the document. Changing the magnitude of the vectors (by multiplying each vector by a constant) has no effect on the angle between them, and hence does not affect cosine similarity. Thus, cosine similarity can be used to compare documents of varying lengths, where term-frequency values might be much higher in one vector compared to the other.

We continue our presentation of background material with a high-level overview of the raw materials which a text-to-space pipeline must start from: language, particularly spatial language.

2.3 Spatial language

There is a wide variety of linguistic tools at an author's disposal to refer to locations and to talk about space[4]. From a linguistic perspective, the building blocks to talk about space and locations include proper nouns (like *London*, *Central Park*, *Paradeplatz*), common nouns (such as *lake*, *city*, *street*), and prepositions (*in*, *at*, *near*, *between*). As language can express seemingly infinite ideas through combining sequences of words according to rules, so too can space be richly described from these same ingredients.

The umbrella term *spatial language* includes not just language about the layout of geographic entities like mountains and cities, but also of smaller manipulable objects, such as coffee cups and computers. Indeed, the literature on spatial language

[4]Discussions of spatial language are centered around the languages processed in this thesis: English and German.

is broad and varied, much of it focused on spatial relationships, particularly through the study of spatial prepositions. Examples are works analyzing which geometric relationships or functional characteristics each preposition specifies or leaves open-ended (see, in particular, Talmy (1983) and Herskovits (1985)) and more applied works building computer systems that attempt to interpret or generate spatial prepositions for some purpose, such as for vehicle navigation systems (Maaß et al., 1995), visually situated dialog systems (Kelleher and Costello, 2008), or image captioning systems (Hall et al., 2015).

In this thesis, the focus is on spatial language about geographic entities such as mountains, cities, and hospitals. These are typically discussed in writing through the use of *placenames*, also known as *toponyms*. Placenames or toponyms are typically defined as the subset of proper nouns which refer to places (Bennett and Agarwal, 2007), where proper noun is a linguistic category for words that refer to particular individuals, such as people and places, and that require capitalization in English and many other languages. Of course, placenames are not the only way to refer to geographic entities or locations: a geographic reference could be made using a place code (such as a postal code or address), a noun ('the city'), a description ('the largest city in Switzerland'), or a complex geographic phrase (such as '20km North of Zürich') (Leidner and Lieberman, 2011). A *geographic reference* is thus any natural language expression referring to a particular geographic entity or region, such as those of a geographical scale (see Montello (1993) for a useful classification of scales). These could also be called *referring expressions* to places, of which placenames are one particularly interesting kind. Some examples of referring expressions to places are shown in Table 2.1.

Table 2.1: Different kinds of referring expressions to places.

referring expression	examples
placename / toponym	Switzerland, Vancouver, Central Park
place code	Winterthurerstrasse 190, Y25L11
compositional description	50 miles south of London
noun phrase with common noun	the city, at the train station
deictic expression	here, over there

2.3.1 Placenames (toponyms)

Placenames are versatile tools to communicate about locations, being used to refer to geographical entities large and small ('Canada' or 'Villa Borghese') and of various types, including administrative regions ('Switzerland'), populated places ('London'), mountains ('Mount Everest'), and water bodies ('Loch Ness'). In the context of text-to-space pipelines, placenames are crucial because, unlike for many of our referring expression examples above, it is usually possible to link particular instances of placenames from a text to specific places and a corresponding geographical representation for these. This process of linking individual toponyms to specific referents / geometries has been termed *toponym resolution* in Leidner (2004a) and is an important step in a text-to-space pipeline, providing the 'glue' around which document-level representations can be built.

While in prototypical examples such as 'London', a toponym or placename is clearly a proper noun and vice-versa, in other cases it is less clear whether a word or a sequence of words is a toponym, and exactly which words are part of a particular toponym. For instance, sometimes an article always accompanies a particular proper noun, such as 'the Netherlands', and often toponyms contain words that refer to a class of geographic entities (feature types), such as 'Park' in 'Central Park'. A related question is whether parts of a placename actually carry semantic content: for instance, 'Lake Placid' is the name of a town as well as a lake. Moncla et al. (2014a) consider the problem of deciding which words constitute a toponym as a kind of ambiguity they call structural ambiguity. Hill (2006) defines a placename slightly differently than a toponym, stating a placename can be composed of a toponym as well as a type (such as 'Ford Hospital') and cautioning that it can be unclear whether the type forms part of the toponym or not.

Delimiting exactly which set of expressions are placenames or toponyms and which are not is a hard problem. According to Coates (2006), proper nouns, including placenames, are those expressions which refer senselessly in an act of communication. Thus, 'properhood' is a mode of reference he calls `PROPER` and contrasts with `SEMANTIC` reference. In `SEMANTIC` reference, the meaning of the

words gets utilized in the act of communication by the speaker and listener. In PROPER reference, the words themselves may carry meaning from an etymological perspective, but in the act of communication this meaning is not considered: names apply or refer *directly* in virtue of an "arbitrary link with what they apply to". In practice, words may have emotional colourings (connotations) in addition to literal meanings (denotations), and it is not always transparent whether an expression is meant in a 'senseless' way or not, as exemplified in Figure 2.1.

Figure 2.1: Strome Ferry: confusion betwen senseless and semantic modes of reference, perhaps due to a broad, touristic audience (there was once a ferry terminal there, but no longer, leading to enough confusion as to warrant *in situ* clarification). Photo credit: Elise Acheson.

While in a person-to-person conversational situation, further clarifications can be obtained from the speaker, this is of course not an option when automatically processing texts. Algorithms must instead rely on contextual clues from the sentence or document, general word usage statistics, and external knowledge about the potential referents such as their population or some proxy for their 'importance'.

2.3.2 Ambiguity and vagueness

Ambiguity and vagueness are both extremely common in spatial language. These two different concepts are not always clearly explained or differentiated, perhaps because

both can apply to the same expression when information is left underspecified. For example, the expression 'We are near London' is ambiguous because London has multiple potential referents (including London, England and London, Ontario, Canada) and is vague because 'near' is highly context-dependent and doesn't have a crisp region of applicability. In general, ambiguity occurs when there are two or more *distinct* possible interpretations for an expression, whereas vagueness applies when there are potentially *infinite* interpretations for an expression, along a continuum of values such as space or time. Examples of vagueness include: Where does a mountain end or begin (Burrough and Frank, 1996)? What distance away does a shop have to be to be 'near' one's home (Worboys, 2001)? Where does 'downtown' begin and end (Montello et al., 2003; Hollenstein and Purves, 2010)?

Ambiguity is extremely common in the context of placenames, since many different places share a name. In the context of parsing text for geographic references, Amitay et al. (2004) define two kinds of ambiguity that need to be resolved: *geo/non-geo* and *geo/geo* ambiguity. Geo/non-geo ambiguity exists whenever a word or expression is both a placename and some other non-geographic reference, such as a reference to a person (such as 'Brooklyn' Beckham vs. 'Brooklyn', New York) or a common word (such as 'nice' vs. 'Nice', France). Geo/geo ambiguity arises whenever a name could refer to more than one place in the world, such as 'Aberdeen, Scotland' or 'Aberdeen, Washington'. The two types of ambiguity can of course also co-exist, as exemplified in Figure 2.2. Toponym (geo/geo) ambiguity has been shown to be spatially autocorrelated at various scales of analysis (Brunner and Purves, 2008) - that is, an ambiguous toponym is likely to have different potential referents in rather close proximity as opposed to randomly distributed in space. This is because toponyms themselves do not arise randomly but rather correlate with language, culture, and the physical landscape (Burenhult and Levinson, 2008). Toponym spatial autocorrelation may complicate disambiguation strategies based on the referents' locations.

Smith and Mann (2003) emphasize the bi-directional nature of ambiguity. Not only can a word or expression refer to multiple geographic entities, but a single

Figure 2.2: Sandwich: geo/non-geo and geo/geo ambiguity co-existing in one placename. Photo credit: Ben Williams.

entity can be known by multiple names, such as 'NYC' and 'the Big Apple' for New York City. These alternate names can be not only abbreviations and colloquial names, but also translations or transliterations when dealing with multiple languages. Axelrod (2003) describe many of these complexities in the context of building a large geographic database containing tens of millions of entries. Their approach involves separating geographic names from geographic entities and associating them to each other using a relational approach, allowing for one-to-many and many-to-one relationships. It is worth noting that ambiguity in the geospatial context is not limited to matching names to particular entities. Another kind of ambiguity could arise when choosing between distinct interpretations for the boundaries of a geopolitical entity (such as in territorial disputes), which Leidner (2007) calls *discord*.

In practice, toponym ambiguity depends on which gazetteer resource provides results for a particular string of text, usually as a set of discrete records. In one gazetteer the string 'San Jose' may have only a single distinct record, whereas in a gazetteer with more detailed (depth) or wider (breadth) spatial coverage, there

could be hundreds of records matching that string. Hence, gazetteers play a key role in the text-to-space process and we discuss them next.

2.4 Gazetteers

Gazetteers are resources which should minimally contain name, geometry, and type information for a set of places (Hill, 2000). They are thus important in linking natural language text (placenames) to geographical space, since they provide geometric representations (such as points, bounding boxes, or polygons) for named places. In addition, gazetteers can be used to identify which words in text are placenames, as well as to obtain information about a place such as population or hierarchical information, which can help in the placename disambiguation process.

2.4.1 Gazetteer production and quality

Gazetteers provide placename information for a defined region of interest, often a country. Indeed, gazetteers have traditionally been produced in a top-down process by national mapping agencies, according to a regulated process and including quality standards and controls. For example, in Switzerland, swissNAMES3D is the official collection of Swiss placenames (or geographic names) produced by the Federal Office of Topography. Accordingly, it adheres to well-defined quality standards, including "full national coverage in homogeneous form and quality" and horizontal and vertical accuracy ranging between 0.2m to 3m depending on the feature type[5].

Increasingly, gazetteers are also being produced using methods ranging from purely bottom-up processes, for example by mining placenames and associated information from crowdsourced data, to processes which integrate a variety of data sources, including authoritative datasets and bottom-up data. The Gazetiki project is one example of a purely bottom-up approach to building a gazetteer, where data from Wikipedia and Panoramio (a discontinued photo-sharing website) were extracted and analysed to automatically generate a set of gazetteer records

[5] As described on https://shop.swisstopo.admin.ch/en/products/landscape/names3D (accessed in 06.2019)

(Popescu et al., 2008). Gao et al. (2017) make use of user-tagged photographs from Flickr, which feature many relevant tags such as 'park', 'museum', and 'river', for their framework to create new gazetteer entries in a scalable and efficient way. More common are the approaches which make use of both authoritative data and data contributed by individuals. Examples of this approach include two very successful projects with global coverage: OpenStreetMap[6] (OSM) and GeoNames[7]. OpenStreetMap is built largely from individual edits by its users but also includes imported datasets such as road networks[8], whereas GeoNames relies heavily on imported authoritative datasets such as ones produced by national mapping agencies and available as open data, but also provides a platform for individual users to contribute data[9].

Because of these varied production processes, data quality in these global datasets varies non-randomly over space, most basically as a function of the data quality of any integrated datasets, and in more complex ways as a property of where and how individual users contribute data. A seminal paper by Haklay (2010) evaluated the data quality of OSM by focusing on positional accuracy and completeness of the street network, showing that there were geographical biases towards more urban and more affluent regions in England. Several works have examined the contents of GeoNames in a particular region or country, including Smart et al. (2010) who mapped and compared the contents of GeoNames in Great Britain with national mapping agency data and crowdsourced datasets, Ahlers (2013) who examined data quality in GeoNames for populated places in Central America, Germany, and Norway, and De Sabbata and Acheson (2016) who compared the contents of GeoNames and the Getty Thesaurus of Geographic Names[10] (TGN) in Great Britain.

[6]https://www.openstreetmap.org/ (accessed in 07.2019)

[7]https://www.geonames.org/ (accessed in 07.2019)

[8]OSM sources are discussed here for example: https://www.openstreetmap.org/copyright (accessed in 07.2019), summarized as "Our contributors are thousands of individuals. We also include openly-licensed data from national mapping agencies and other sources (...)"

[9]GeoNames sources are discussed here for example: https://www.geonames.org/about.html (accessed in 07.2019)

[10]http://www.getty.edu/research/tools/vocabularies/tgn/

What is clear from these studies is that the contents of gazetteer resources in general can vary greatly from one resource to another, and thus tasks making use of a gazetteer, for example in a placename search task, are likely to lead to different outcomes depending on which resource is chosen. Most gazetteers, including OSM and GeoNames, can be queried via service-oriented architectures or APIs[11]. Hence their contents can be relatively easily accessed by a text-to-space pipeline to obtain place candidates for a particular placename or location string, including a geometry (usually a point coordinate), and GeoNames in particular is heavily used for this purpose in the academic literature (e.g. Van Laere et al., 2014; Weissenbacher et al., 2015; van Erp et al., 2015; Karimzadeh et al., 2018).

2.4.2 Gazetteer matching and integration

A second related stream of research focuses on integrating gazetteers and enriching existing gazetteers, for example by adding more records or by enriching the annotation of existing records (Hastings, 2008; Smart et al., 2010; Gelernter et al., 2013). As a particular geographical area of interest may be covered by different resources, with no full agreement on the place names, types, or geometries present across the common region, it may be desirable to form a single, integrated, duplicate-free gazetteer resource in order to facilitate and improve performance on a range of tasks. A key part of this process is to establish links between records that are deemed to be about the same real-world entity. This step is generally called record linking, or in the context of linking gazetteer records, *gazetteer matching* (Acheson et al., 2019).

Gazetteer matching is typically performed by comparing records based on their names, geometries, and optionally feature types, with a decision on each potentially matching record pair reached either using hand-crafted rules (Fu et al., 2005; Hastings, 2008; Smart et al., 2010; McKenzie et al., 2014) or using machine learning (Sehgal et al., 2006; Zheng et al., 2010; Martins, 2011; Gonçalves, 2012). Many works have focused on populated places and points of interest (Martins, 2011; Zheng et al.,

[11]For example, GeoNames directly provides a set of web services, and a variety of options exist to query OSM, including one search (geocoding) service known as OSM Nominatim.

2010; Dalvi et al., 2014; McKenzie et al., 2014), and only few have worked with a broader mix of feature types (Sehgal et al., 2006; Hastings, 2008; Smart et al., 2010).

Natural features in particular present interesting challenges in the context of gazetteer matching, including vagueness, since feature types like mountains and valleys are classic examples of hard-to-delineate (vague) entities (Burrough and Frank, 1996), and potentially a high degree of name ambiguity (Derungs and Purves, 2013). In addition, records in different gazetteers may be annotated with types from different feature type hierarchies, which means matching methods must potentially align these hierarchies to each other before any matching is attempted (Hastings, 2008; Morana et al., 2014), or potentially use data-driven approaches based on annotated record subsets and their type alignments (Brauner et al., 2007; Sehgal et al., 2006). Feature type alignment is itself a complex problem which receives dedicated research attention (Janowicz and Keßler, 2008; Zhu et al., 2016).

Given the shift in gazetteer production processes, which amalgamate multiple resources into one and integrate user-generated content, and the increasing number of applications operating on multi-country or global scales, which require fit-for-purpose placename resources, there is a need for effective gazetteer matching methods which can readily be transferred to new application contexts.

2.5 Text-to-space methods

A large body of research exists which computes document-level geographic representations for various data sources and for a variety of tasks/purposes. This work is associated with a variety of terms, including document *geographic scope* (Buyukokkten et al., 1999; Ding et al., 2000; Silva et al., 2006; Anastácio et al., 2009), *geographic focus* (Amitay et al., 2004; Lieberman et al., 2007), *geographic footprint* (Markowetz et al., 2005; Purves et al., 2007; Bordogna et al., 2012; Derungs, 2014), *georeferencing* (Martins and Silva, 2005; Van Laere et al., 2013), *geolocation* (Wing and Baldridge, 2011; Dredze et al., 2013; Jurgens et al., 2015; Rahimi et al., 2016), and *locational focus* (Yin et al., 2014). Methodologically, there are fundamentally two different approaches used for this process: *gazetteer-based* approaches and

supervised approaches. Gazetteer-based approaches (or pipelines) form the focus of this thesis and are described in detail later in this section. However, for completeness it is important to give a brief overview of supervised approaches.

In supervised approaches, geometries can be assigned to text documents without the use of gazetteers to obtain geographic representations for named places. Instead, a spatial representation for a document is obtained or inferred based on the document's similarity to other documents which have 'known' (human-assigned) geometric representations - that is, geometries are derived from training data (whence *supervised*). This requires having a large corpus of text documents where each is associated with a particular geometry; common examples are Wikipedia articles manually tagged with latitude-longitude coordinates (Wing and Baldridge, 2011; Dias et al., 2012; Van Laere et al., 2014), and microblog (typically Twitter) posts tagged with GPS coordinates from a user's device (Roller et al., 2012; Dredze et al., 2013; Jurgens et al., 2015). Such a 'geotagged' corpus can then be used to train a language model, or a classification/regression model, which can assign geometries to untagged documents. The dominant approach is to build geographical language models from the training data, which relies on the assumption that similar language characterizes documents that originate from similar locations, and that these associations can be learned. Importantly, language models are built using not just placenames, but potentially the entire textual content of documents, including 'location indicative words' such as regional expressions and colloquial names for places (see for example Han et al. (2014)).

Hence, gazetteer-based approaches and supervised approaches differ fundamentally in how they obtain geometric representations of documents: gazetteer approaches rely on geometries found in gazetteer resources, whereas supervised approaches rely on the 'ground truth' geometries assigned to similar documents. It is worth noting that there is not a strict separation between pipelines using these two approaches, as gazetteer-based pipelines can make use of supervised learning or machine learning, for example to detect placenames in text, and supervised pipelines can avail themselves of gazetteers, for example by prioritizing toponyms

over other words in the text. In the remainder of this thesis, we focus solely on gazetteer-based pipelines.

Gazetteer-based pipelines typically consist of 3 main steps. First, textual references such as placenames are extracted from the document, a step referred to varyingly as *toponym recognition*, *geoparsing*[12], or *georecognition* (Leidner and Lieberman, 2011). Second, in a referent disambiguation step, known as *toponym resolution* (Leidner, 2007), *geocoding*, or *grounding* in the literature, each place reference is linked or 'grounded' to a gazetteer entry including an explicit geographical representation, such as a latitude-longitude coordinate or a polygon. In a final step, an overall geographical representation is determined by further processing the set of grounded locations found for a document. Some steps may overlap, for example if using a preliminary document-level representation (step 3) to aid in disambiguation/grounding (step 2). We now look at these steps in more detail.

2.5.1 Identifying placenames

The first step in gazetteer-based pipelines is to identify words or sequences of words in text that refer to locations - typically toponyms/placenames, but potentially other types of referring expressions to places (Table 2.1). Leidner and Lieberman (2011) identify three main approaches to placename identification: *gazetteer lookup based*, *rule based*, and *machine learning based*. Gazetteer lookup methods involve going through text word-by-word looking for string name matches in a gazetteer. Since many common words, like 'bath', 'nice', and 'of', also appear in gazetteers as placenames, gazetteer lookup methods should ideally be accompanied by pre-processing (selecting a subset of words from the text to look up, for example proper nouns) and post-processing (such as removing common false positives via an exclusion list) to deal with geo/non-geo ambiguity. A gazetteer-based approach is used in Fisher et al. (2011) to detect locations in scientific articles, where only word sequences of up to 3 words taken from the title, keywords, and abstract are

[12]Confusingly, the term 'geoparsing' is now also being used to refer to the combination of identifying and grounding placenames, i.e. steps 1 and 2 together (e.g. Gritta et al., 2017; Karimzadeh et al., 2018)

looked up, with words in the abstract limited to those starting with a capital letter, a suitable strategy for formal English-language text. In Tamames and de Lorenzo (2010), a part-of-speech (POS) tagger is first run over the text, and only capitalized words within noun phrases are looked up in a gazetteer.

Rule based methods look for place references based on rules or patterns, including sequences of words or common toponym structures. In Leveling (2015), a rule based approach is used to detect geographic and geologic locations in tables and captions of scientific articles. Examples of patterns detected include sequences containing capitalized words, words appearing in a gazetteer, location modifiers (e.g 'north-west', 'central'), and feature types (e.g. 'river', 'mountain', 'shield').

Machine learning based methods make use of annotated 'gold standard' corpora to find statistical associations between place references and various predictive features derived from the text. Examples of features include whether a word is capitalized, what the POS tag of a word is, whether a word is preceded by a word like 'in' or 'to', whether a word is the first word in a sentence, and so on. These associations are then used to identify locations in previously unseen text. Most NER tools are machine learning based, including the widely-used *Stanford NER* tool (Finkel et al., 2005) which has been shown to be high-performing over a range of datasets (Jiang et al., 2016). NER is often considered an overarching problem to location/placename identification, dealing with a wider categories of named entities than just locations. Indeed, a wide variety of NER tools are available, which vary in their categorization schemes. The 'location' entity is one of the most studied types, alongside 'person' and 'organization' (Nadeau and Sekine, 2007), and these 3 categories are the ones used in Stanford NER's 3-class model. In contrast, one of the NER categorization schemes in the NLP library *spaCy* features a wider range of categories originating from the OntoNotes 5 corpus[13], with locations potentially being classified as 'location' (LOC), 'facility' (FAC), 'organization' (ORG), or 'geo-political entity' (GPE), as shown in Figure 2.3.

[13]https://catalog.ldc.upenn.edu/LDC2013T19 (accessed in 07.2019)

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.

Figure 2.3: Entity types in one of spaCy's categorization schemes for their NER module.

In addition to categorization schemes, NER tools also vary in terms of the algorithms used and the text type or genre (such as news, web pages, or social media)[14]. In general, NER tools perform best on testing data that closely resembles the training data used to train the model (Augenstein et al., 2017), given sufficient training data, a situation analogous to most (machine) learning problems. Most widely used, general-purpose NER tools are trained on formal text, including news articles, and hence informal texts such as tweets are challenging for standard NER tools and strategies (Liu et al., 2014). One solution is to re-train a model on suitable data, as in Yin et al. (2014) who re-train Stanford NER on Twitter data in order to identify location mentions in the content of tweets.

Mixing placename identification methods is also possible, where different strategies may complement each other and make up for their respective weaknesses.

[14] A useful note should be mentioned here regarding text *genre* and text *domain*, from Augenstein et al. (2017): "These are two dimensions by which a document or corpus can be described. Genre here accounts the general characteristics of the text, measurable with things like register, tone, reading ease, sentence length, vocabulary and so on. Domain describes the dominant subject matter of text, which might give specialised vocabulary or specific, unusual word senses. (...) One notable exception to this terminology is social media, which tends to be a blend of myriad domains and genres, with huge variation in both these dimensions (...)"

Gelernter and Zhang (2013) build a Twitter-specific location parser which combines the results of four parsers, as they found individual out-of-the-box tools struggled with abbreviations, misspellings, fine-grained places, and colloquial names in tweets. Won et al. (2018) combine 5 NER systems using a voting system in order to increase placename identification performance on challenging historical text.

2.5.2 Grounding placenames

The second step in gazetteer-based pipelines is to *ground* the placenames or location strings identified in the first step, which means linking each location unit to a unique gazetteer record and its associated geographical footprint (such as a point, bounding box, or polygon). This step is also known as *toponym resolution*, a term introduced in Leidner (2004a). Yet another term for it is *geocoding*, a term whose meaning has evolved as geocoding tools themselves evolved from taking structured addresses as input and returning geographic coordinates, to taking a wide range of location strings as input and potentially returning richer geometries.

The main challenge in this step is to deal with geo/geo ambiguity: finding the correct intended referent location for each location string. A systematic analysis of disambiguation strategies is presented in Leidner (2007). Buscaldi (2011) partitions approaches to disambiguating placenames into: *map-based*, *knowledge-based*, and *data-driven or supervised* approaches. Map-based methods rely on geographic properties such as proximity of resolved toponyms to each other, while knowledge-based methods make use of external information such a population data, and data-driven methods base disambiguation on statistics from annotated corpora.

Specific disambiguation strategies in the literature are varied, can be used in combination, and include:

- Looking for a country or containing place within the sentence or a certain word span from the textual place reference (Amitay et al., 2004).
- Assigning the placename to the referent with the highest population as a default (Amitay et al., 2004).

- Assigning the placename to the 'most important' referent as a default, where importance could be calculated using place hierarchy, feature types, or other information available in a gazetteer (Clough, 2005).
- Using a 'one sense per discourse' heuristic, meaning, always linking a particular placename to the same referent within a text document (or 'discourse') (Leidner, 2007).
- Using usage statistics based on large-scale text mining, where 'importance' of a specific referent could be how often that referent appears in text (Rauch et al., 2003).
- Using placename co-occurrence information (e.g. crawled from Wikipedia) to disambiguate based on textual context (Overell and Rüger, 2008).
- Picking the referent which minimizes the distance to all other already-disambiguated placenames Smith and Crane (2001).
- Picking the referent which minimizes the maximum or total distance between all referent candidates for a document (Moncla et al., 2014a).
- Looking for type information in the textual content of the placename (Batista et al., 2010).

The potential for ambiguity typically increases along with the depth and breadth of the gazetteer(s) used. For example, Clough (2005) calculated that about 8% of placenames in the UK-specific Ordnance Survey had multiple entries, compared to 59% of names in the global-coverage TGN.

2.5.3 Geographically representing text documents

The two previous processing steps, identifying placenames and grounding placenames, should result in a set of placenames alongside a geometry for each. This output can be considered the output of the toponym resolution process, concerned with individual mentions of toponyms and their geographical representation, typically a latitude-longitude point for each. The third and final step seeks to generate a useful representation for not just individual toponyms or placenames, but for the document as a whole, based on the task/purpose. This step is a defining feature of works

aiming to go further than identifying and resolving place references in text. Not performing this step results in a 'bag of locations' representation for a document, and if geometries are points, a 'bag of points' representation. This extra step's goal is to summarize in some useful way the grounded place references obtained in a document, such as by emphasizing or picking out key locations over others.

Indeed, examples of what can be done in this third step are to select one or several 'important' places among a wider set, and/or to create a new geometry for a document out of, for example, a set of points (such as a bounding box or polygon containing all points, or a subset of more 'important' points). Correspondingly, geographical representations for documents in the literature tend to vary in terms of how many locations are retained in the model (such as one, several, or all grounded place references), and what footprint type is used (such as a set of points, a bounding box, or a grid).

Table 2.2 presents footprint types commonly used in gazetteer entries (Hill, 2000), and thus also the options typically used to represent individual grounded placenames. Other representational options for individual entities are possible: in Axelrod (2003), 6 types of spatial representation for geographic entities are presented and used in a geospatial database, including an area 'without clearly defined boundaries' and a probability density distribution.

type of representation	description
point	single pair of latitude & longitude coordinates
bounding box	double pair of coordinates representing the maximum and minimum of latitude and longitude extent
line	set of points that do not enclose a space
polygon	set of points that do enclose a space
grid representation	grid references to a location according to an identified grid referencing scheme

Table 2.2: Types of geospatial representation in Hill (2000).

As for ways of geographically representing a text document, options include using placenames as-is, using a single point, using a set of points, using a polygon

or bounding box, using a grid, and using a probability density function. These document-level options are illustrated in Figure 2.4. For applications performing spatial reasoning, at least some of the geometries used should be areas (such as bounding boxes or polygons) rather than points (Vasardani et al., 2013) so that footprints can be compared using containment relationships or overlap. Indeed, limited spatial reasoning is possible when dealing only with points, being largely limited to the point-to-point distance, though deriving areas from points is one useful strategy to create richer geometries (Galton and Duckham, 2006).

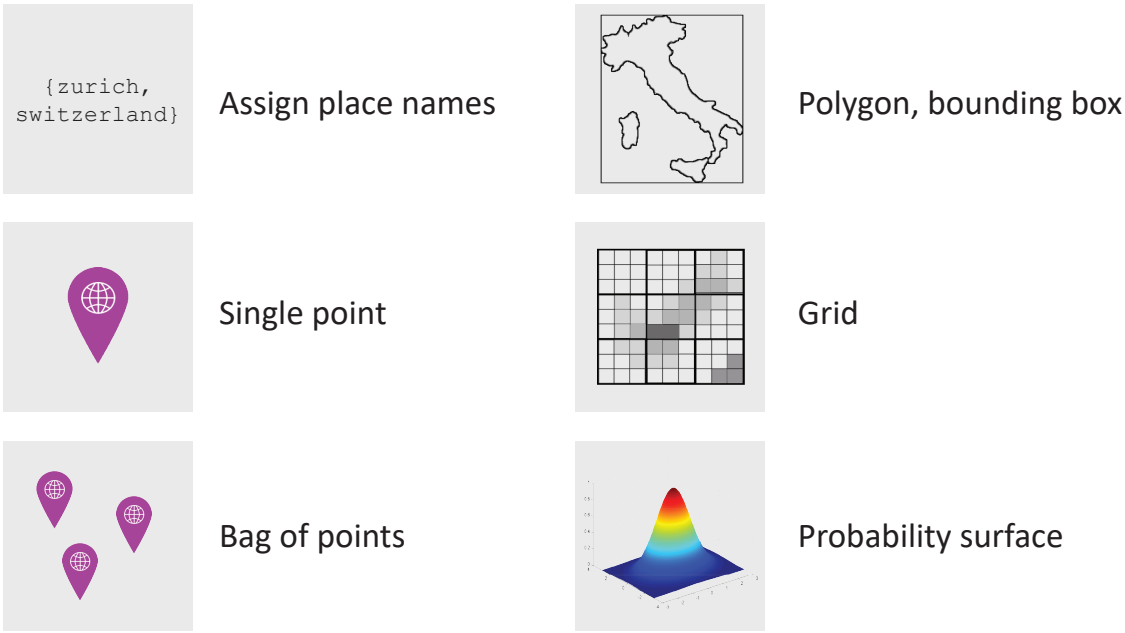


Figure 2.4: Different possible geographical representations for a text document.

As mentioned, document-level geographic representations are often called the 'geographic scope' or 'focus' of a text document (Buyukokkten et al., 1999; Amitay et al., 2004), representing the geographical locations/region the document is (mostly) about. Many early works identified geographic scopes for webpages in order to enhance traditional search engines by returning geographically relevant results (Buyukokkten et al., 1999; Ding et al., 2000; Amitay et al., 2004; Markowetz et al., 2005; Wang et al., 2005; Zong et al., 2005). Below, significant early works on the problem of assigning geographic representations to documents are briefly

presented, including their representational choices, and some information on their methods or objectives, where relevant:

- Woodruff and Plaunt (1994): Their *GISPY* spatial indexing system represents each place found in a text as a polygon, which are then overlayed, and each document gets assigned zero or more polygons based on their weights in the overlay.
- Buyukokkten et al. (1999): They model each location that has links to a webpage of interest as a circle, where the radius of the circle represents a normalized count of links; these circles are displayed on a map for visualization.
- Amitay et al. (2004): Their *Web-a-where* system assigns up to four locations as the 'geographic focus' of a web page. Locations considered distinct enough from each other and surpassing some importance threshold are retained as the geographic model of the document, with each location tied to a node in their hierarchical gazetteer providing a relevant geometry (point or polygon).
- Markowetz et al. (2005): They use a grid representation for their document footprints, partitioning the total geographical area of interest into equal area tiles, then they assign an integer value to each tile representing 'the certainty that the document is relevant to the tile'.
- Purves et al. (2007): The *SPIRIT* GIR system assigns footprints to documents based on the locations found using gazetteer lookup, with a gazetteer containing polygonal footprints for each location. However, for processing efficiency, locations in documents are represented using bounding boxes derived from these polygons, and ultimately document footprints become part of a grid-based spatial index.
- Lieberman et al. (2007): Their *STEWARD* system retains all locations found in a text document, ranked according to a focus score. Their scoring algorithm considers the mention count of a location, as well as geographic proximity and textual distance between pairs of locations. Each location is tied to a gazetteer entry containing latitude-longitude coordinates, and no attempt is made at jointly modeling these points. In a follow-up work focusing

on news articles (Teitler et al., 2008), some domain-specific refinements to their processing pipeline are made, including assigning greater importance to locations mentioned earlier in text.

The works presented above are but a small subset of a much larger body of literature aiming to assign geographic representations to documents for various purposes. Text-to-space pipelines and geographic representation choices vary from one work to the next, with early local search applications generally associating documents with one or more key places, GIR systems assigning potentially richer spatial footprints to each document for indexing and retrieval, and many systems modeling locations at all levels of granularity using simple latitude-longitude coordinates. Methods and modeling choices are often tightly coupled to the underlying data and resources, the tools used, and the intended application, though assumptions made and possible alternatives are not always stated.

2.6 Research gaps

Based on the overview of relevant background and literature presented above, the following research gaps have been identified:

Spatial properties of global gazetteer resources. Several studies have looked at data quality in widely used, global spatial resources such as GeoNames, TGN, and OpenStreetMap, but typically focusing on just one resource in isolation and on a particular geographic region. GeoNames in particular is widely used in text-to-space pipelines but its global spatial properties, likely to influence the spatial properties of task results, have not been sufficiently examined. Some studies have found that its coverage is unbalanced in particular countries, but few in-depth systematic global analyses have been conducted. In particular, few works have studied coverage across country boundaries and have looked at a wider range of feature types than populated places, including natural feature types such as mountains, hills, and streams. Methodologically, a global quality analysis is not straightforward because there is no accepted, high quality authoritative resource to use as a comparison.

Gazetteer matching methodology. Many gazetteer matching or integration studies exist, but it is unclear what methods should be applied to a new gazetteer matching task and how to choose a matching strategy. Indeed, each work on this problem differs in terms of what data is used, what methods are implemented, what matching features are used, and how the task is evaluated. Gold standard datasets and implementation code are nearly absent from the publication landscape. As with gazetteer coverage studies, few gazetteer matching studies have looked at a wider range of feature types, including natural features such as mountains, hills, and streams, which may present additional challenges due to vague boundaries and differing gazetteer representations. It is also unclear how to integrate feature type information into the matching task, particularly when dealing with more than one feature type hierarchy.

Natural feature types. Most text-to-space research, including work on gazetteer resources, focuses on a common subset of feature types such as populated places (cities, towns, villages) and geopolitical entities like states and countries. Natural feature types such as mountains, hills, and streams, receive comparatively little attention despite presenting interesting research challenges. Dealing with texts that contain placenames for natural features requires having suitable, potentially integrated, gazetteer resources and a strategy for recognizing and disambiguating these placenames when many tools focus on populated places.

Textual data sources. Early work on assigning geographic representations to documents focused on webpages for geographical search, and since then, the two dominant types of documents fed to text-to-space pipelines have arguably been news articles and Twitter content. Other types of documents, such as blogs, reports, scientific articles, historical text, and hiking descriptions, receive limited attention but present their own set of challenges. These challenges include a shift to dealing with fine-grained places and natural features, as in parsing hiking descriptions (Moncla et al., 2014a), and working with longer documents where only a subset of locations appearing in the text are relevant for a geographical model, as with

scientific articles (Fisher et al., 2011). Generally, different text genres and text domains may vary in terms of which subsets of feature types (such as natural features, urban POIs, or populated places) and referring expressions to places (such as toponyms or compositional descriptions) (Table 2.1) appear most often.

Geographical representation. Most text-to-space pipelines in the literature output a single point or a set of points to represent a text document, often with no further processing of the geometries output at the toponym resolution step. However, even when working with a gazetteer resource that returns points for all named places, more complex geometries can be created from a set of points (Galton and Duckham, 2006), such as a bounding box or a convex hull. Few works explore these representational options at all.

Real-world motivated case studies. There is a tendency in the literature to treat a text-to-space pipeline as an end in itself, rather than truly as a means to an end. Though it is interesting to advance research based on datasets that present interesting properties, it is also important to pursue avenues of research that are ultimately motivated by real-world needs.

Reproducibility. Text-to-space pipelines involve not just specific tools and specific snapshots of gazetteer resources, but also specific computer code that implements the particulars of placename detection, placename grounding, and the geographical representation of documents. Despite the importance of implementation details on any results, publications which provide access to the code and data used to generate these results are all too rare. The reproducible research movement promotes the idea that, as scientific analyses become increasingly computational, code and ideally data should be made freely available alongside publications in order for others to be able to more easily verify any results and build upon them (Peng, 2011). Reproducibility also means that the highly complex, decentralized enterprise that is modern scientific research can potentially become more cooperative, and less adversarial, and ultimately provide more value to the public.

Research contributions relating to text-to-space pipelines are varied and have come from computational linguists, who focus on improving NER and other NLP tools, computer scientists, whose skills includes building specific applications that are scalable and efficient, and GIScientists, who can apply their expertise to areas of improvement relating to spatial resources and models. There are many ways to further advance text-to-space research, and as a GIScientist, the interdisciplinary work presented in this thesis maintains a focus on geographical space throughout.

There is remarkable diversity in approaches to the description of geographic places and, until recently, no standardization beyond authoritative sources for the geographic names themselves.

— Linda Hill, *Georeferencing* p.94

3

Text-to-space resources: gazetteers

Contents

The contents of this chapter are based on two publications included in this thesis: Paper I (Acheson et al., 2017a) and Paper II (Acheson et al., 2019).

This chapter concerns the first theme of this thesis: text-to-space *resources* known as gazetteers. Gazetteers play an important role in many text-to-space pipelines, and provide the crucial link from textual placenames to geographical representations in gazetteer-based pipelines. The first part of the chapter (section 3.1) presents work done to quantitatively analyze and compare spatial properties of two global gazetteers: GeoNames and the Getty Thesaurus of Geographic Names (TGN). The second part of the chapter (section 3.2) then presents work on gazetteer matching, where records from GeoNames are linked to records in swissNAMES3D when a pair of records is a 'match', that is, when the records represent the same real-world entity.

3.1 Gazetteer comparison and analysis

The spatial properties of two widely-used global gazetteers, GeoNames and TGN, are examined in this section. Previous work included single-gazetteer studies of GeoNames, such as an analysis of data quality and spatial coverage for specific countries and regions (Ahlers, 2013), and a global study mapping the density of populated places in the resource and comparing it to population data (Graham and Sabbata, 2015). In a preliminary work developing our methods to compare the spatial distribution of records across gazetteers, we focused on Great Britain and compared data in GeoNames and TGN to data in two authoritative gazetteers by Ordnance Survey: OS 50k and its more recent replacement, OpenNames (De Sabbata and Acheson, 2016). The follow-up work presented here (Acheson et al., 2017a) analyzes not just one country or region, but the global spatial coverage of both GeoNames and TGN, and not one feature type, but a range of common types, including populated places and natural feature types. Our gazetteer data, methods of analysis, and results are detailed below.

3.1.1 Gazetteers

GeoNames and TGN are briefly described below, followed by a short introduction to gazetteer data quality assessment.

GeoNames

GeoNames is a widely used placename resource, commonly used in text-to-space pipelines, particularly during the grounding/disambiguation step (e.g. Van Laere et al., 2014; Weissenbacher et al., 2015; van Erp et al., 2015; Karimzadeh et al., 2018). Its many appealing properties include global coverage, a very large number of records (over 10 million records already in June 2015), and daily data exports which can be freely downloaded and used. The records in GeoNames come from a variety of sources, including authoritative datasets and individual contributors. GeoNames provides for each record the standard elements of name, type, and geometry, and for many records offers additional information such as alternate names, population

information, and information about containing countries or regions (hierarchical information). Each record is represented by a latitude-longitude point in the free version, including point centroids for records like countries and lakes. Records are assigned a feature type from a two-level feature type hierarchy: one of 9 feature classes and one of 645 feature codes (each a subdivision of a feature class). Despite a large number of available feature codes, the distribution of these codes in the resource is highly skewed, with around a third of all records in the gazetteer having the code 'PPL' for 'populated places' (Figure 3.1).

TGN

TGN is a gazetteer with an explicit focus on places of cultural and historical significance and a target audience which includes museums, archivists, and researchers in art and art history. As stated in their own words, "TGN is intended to aid cataloging, research, and discovery of art historical, archaeological, and other scholarly information. However, its unique thesaural structure and emphasis on historical places make it useful for other disciplines in the broader Linked Open Data cloud.[1]" The dataset has been accessible as Linked Open Data since 2015. It is a curated resource with over 1.4 million records of named places, including political entities such as cities and countries, and natural features such as mountains and streams. Records are annotated with a name, feature type (potentially a primary 'preferred' type alongside other secondary types), latitude-longitude coordinates, hierarchical information, and potentially many other fields such as a descriptive note, alternate names, related places, and historical relationships. TGN has been used in the context of text-to-space pipelines (e.g. Smith and Mann, 2003; Clough, 2005; Overell and Rüger, 2008). TGN's feature types (or 'place types') are from the controlled vocabulary of Getty's own Art & Architecture Thesaurus[2]. As with GeoNames, the distribution of records in the resource is also skewed to a small number of types, including the 'inhabited places' type (Figure 3.1).

[1]From http://www.getty.edu/research/tools/vocabularies/tgn/about.html (page version from 21.03.2019)

[2]Available online at http://www.getty.edu/research/tools/vocabularies/aat (accessed in 07.2019)

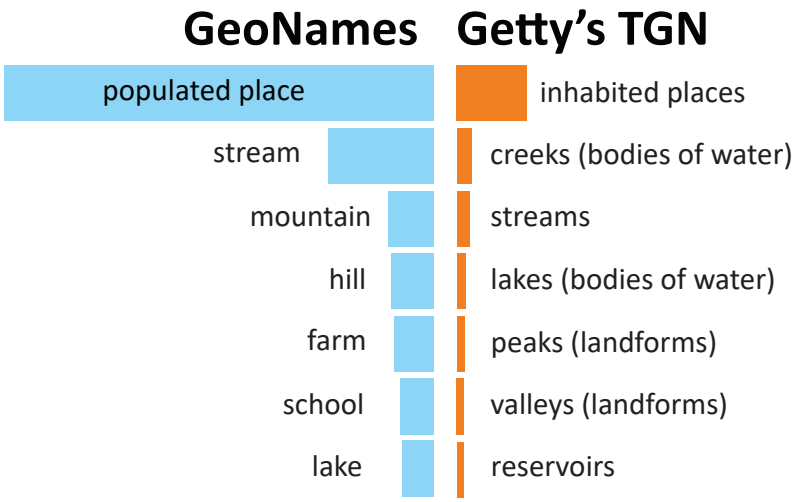


Figure 3.1: Most frequent features types for GeoNames and TGN. Figure adapted from Acheson et al. (2017a).

Gazetteer quality

Evaluating gazetteer quality can be done along many dimensions or components, including the five major dimensions from the US Federal Geographic Data Committee to assess geospatial data quality: attribute accuracy, positional accuracy, logical consistency, completeness, and lineage (Guptill and Morrison, 1995). Relating to gazetteers specifically, Leidner (2004b) lists seven criteria for 'gazetteer selection', which Hill (2006) adapts and expands into the 8 gazetteer quality criteria presented in Table 3.1.

In Table 3.2, we take this list of 8 criteria and apply it to our gazetteers of interest, GeoNames and TGN, alongside two authoritative gazetteers by national mapping agencies: OS 50k for Great Britain, and swissNAMES3D for Switzerland. This preliminary evaluation along a set of defined dimensions helps illustrate some important differences between our two global gazetteers of interest and the authoritative resources: GeoNames and TGN contents have a more varied lineage (sources), their precision is not well defined, and their completeness and balance are both essentially unknown. Indeed, completeness is an issue because some regions of the world where little data is available will almost certainly be poorly covered, and this means balance is also affected, since different regions will not be covered with

Table 3.1: Gazetteer quality criteria from Hill (2006). Table from Acheson et al. (2017a).

criterion	description
availability	"Degree to which the gazetteer is freely available and not limited by restrictive conditions of use"
scope	"Small communal database, regional/national coverage, or worldwide coverage"
completeness	"Degree to which the scope of the gazetteer is covered completely"
currency	"Degree to which the gazetteer has incorporated changes"
accuracy	"Number of detectable errors in names, footprints, and types"
granularity	"Includes large, well-known features only or features of all sizes and those that are less well known"
balance	"Uniform degree of detail, currency, accuracy, and granularity across scope of coverage"
richness of annotation	"Amount and detail of descriptive information, beyond the basics of name, footprint, and type"

a 'uniform degree of detail, currency, accuracy, and granularity'. Data quality in general may vary in ways related to data availability for a particular region.

Table 3.2: Gazetteer quality criteria evaluated for four gazetteers including GeoNames and TGN. Table adapted from Acheson et al. (2017a).

criterion	GeoNames	TGN	OS 50k	swissNAMES3D
availability	free	free	free	free
scope	worldwide	worldwide	Great Britain	Switzerland
completeness	?	?	✓	✓
currency	daily	two weeks	annual	annual
precision	varied	approximate	1k grid cell	0.2m–3m
granularity	medium to fine	medium	medium to fine	fine
balance	?	?	uniform	uniform
lineage	various sources	GNIS, experts	OS maps	SwissTopo maps
richness of annotation	medium	rich for portion	medium	medium

Since there is no global, authoritative 'ground truth' resource to compare GeoNames and TGN to, we must proceed in other ways in order to draw conclusions about quality criteria such as completeness and balance. One way is to quantify and visualize the *coverage* of the resources, a property not explicitly listed in Tables 3.1

and 3.2 as a quality criterion but related to scope, completeness, granularity, and balance. We define coverage of a gazetteer resource as the feature density across space ('spatial coverage', as in Hill (2006), p. 144). Balance in particular, as defined in Table 3.1, depends on this feature density across space - specifically, how uniform the coverage is. We focus on this particular aspect of balance, rather than balance along other dimensions such as accuracy and richness of annotation.

3.1.2 Methods

We obtained full data snapshots of GeoNames and TGN on 30.06.2015. In addition to analyzing the full contents of both gazetteers, we selected a subset of feature types to analyze in isolation, starting from the types with the most records in GeoNames: populated places, streams, mountains, and hills. We chose corresponding types in TGN by considering type names, definitions, hierarchies, and record counts. The types selected for analysis in each resource, including counts and percentage of total records, are presented in Table 3.3.

Table 3.3: Feature types selected for analysis in GeoNames and TGN and their counts and percentage of total records in each resource. A * indicates that any record with a feature code matching this base was included in the count. Table adapted from Acheson et al. (2017a).

feature type		GeoNames			TGN	
name	codes	count	%	place types	count	%
all	(all)	10,203,772	100.0	(all)	1,439,138	100.0
populated places	PPL*	3,767,721	36.9	inhabited places	564,533	39.2
streams	STM*	1,027,913	10.1	streams	108,445	7.5
mountains	MT*	390,418	3.8	mountains	25,975	1.8
hills	HLL*	351,926	3.4	hills	25,680	1.8

We analyzed the global coverage of records in GeoNames and TGN by aggregating records using 3 different units: fine-grained, equal area 10x10km cells; medium-grained, equal area 100x100km cells; and coarse-grained, varying-sized country units. Using the fine-grained units, we produced global coverage maps using the ArcGIS Point Density tool (10x10km cells with 30km neighborhoods) in the equal area Goode Homolosine Land projection, for the 5 data subsets (including all

features) presented in Table 3.3. These 10 maps allow for a visual overview of coverage in each resource for each data subset, as well as quick insight into how coverage varies across resources in different parts of the world and across national boundaries. Using the medium- and coarse-grained units, we compared counts in corresponding units across the two gazetteers using Kendall's tau rank correlation (for non-normally distributed data). Finally, using again these two coarser-grained units, we established descriptive, quantitative relationships between counts in the two datasets using linear models, which use not just rank but also magnitude.

3.1.3 Results and interpretation

We first mapped the global coverage of GeoNames and TGN for all records, rendered in Figure 3.2 in terms of GeoNames quantiles for record density. First of all, the TGN map shows visibly lower density nearly everywhere compared to the GeoNames map, which is to be expected since TGN contains only about a tenth of the total number of records that GeoNames has (Table 3.3). Secondly, in both maps one can see that coverage varies widely in different parts of the world, with generally denser coverage in North America and Europe and sparser coverage in South America and Africa, a pattern which is especially pronounced for TGN. Thirdly, the country unit seems to be driving some sharp changes in coverage in both maps: in TGN for example, Germany seems very densely covered compared to the surrounding countries, especially to the East, and in GeoNames, coverage seems to suddenly drop when crossing the border from Norway to Sweden. Hence we may be seeing some effects of the integration of (open) datasets from national mapping agencies.

Figure 3.3 shows the global coverage of GeoNames and TGN for the 4 feature type subsets: populated places, streams, mountains, and hills. We see similar patterns of coverage in these maps as in the maps of all records, with generally sparser coverage in TGN compared to GeoNames, and overall coverage getting both sparser and more idiosyncratic in both resources as the overall number of records gets smaller (from populated places down to hills, c.f. Table 3.3). The map of hills in TGN, for example, shows just how few records are catalogued at all for this

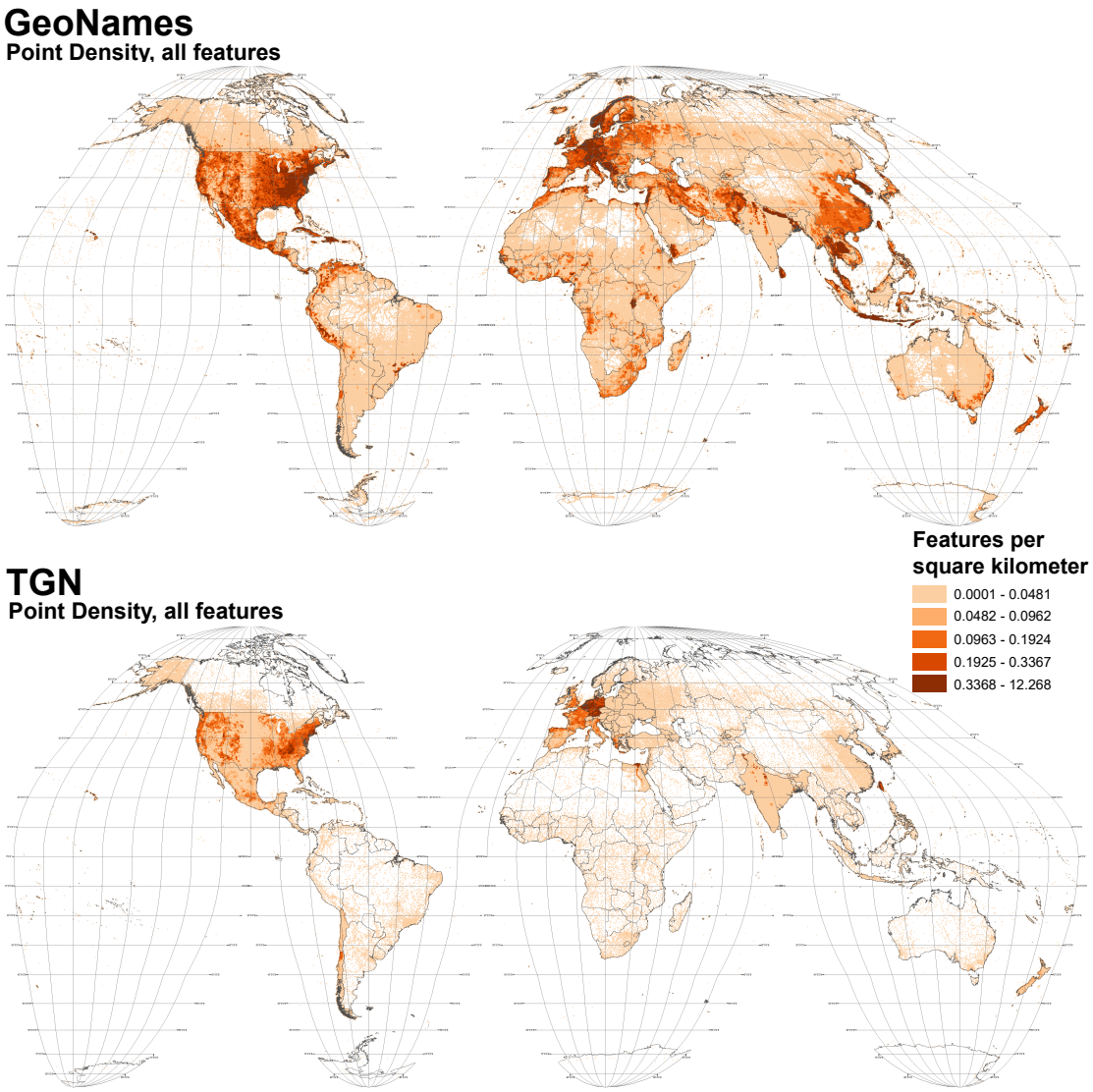


Figure 3.2: Point density maps for all features in GeoNames (top) and TGN (bottom), rendered in terms of GeoNames quantiles, in the Goode Homolosine Land projection. Figure from Acheson et al. (2017a).

feature type, with most of the world map showing a density of 0, and a sudden increase in density in a particular country, Germany. Thus, we again see the country unit as a driver of coverage, not just overall but also for specific feature type subsets.

The Kendall's tau rank correlations for corresponding 100x100km cells and country units in GeoNames and TGN are presented in Table 3.4[3]. In general,

[3]Note that for the country units, a pair of counts was included in the rank correlation calculation when neither count was 0, whereas for the 100x100km cells, much more numerous than the number of countries, a pair of counts was included in the calculation as long as one of the counts was not 0. This avoids artificially high correlations due to matching counts of 0 in the country counts, but also avoids dropping too many meaningful pairs for the raster cells.

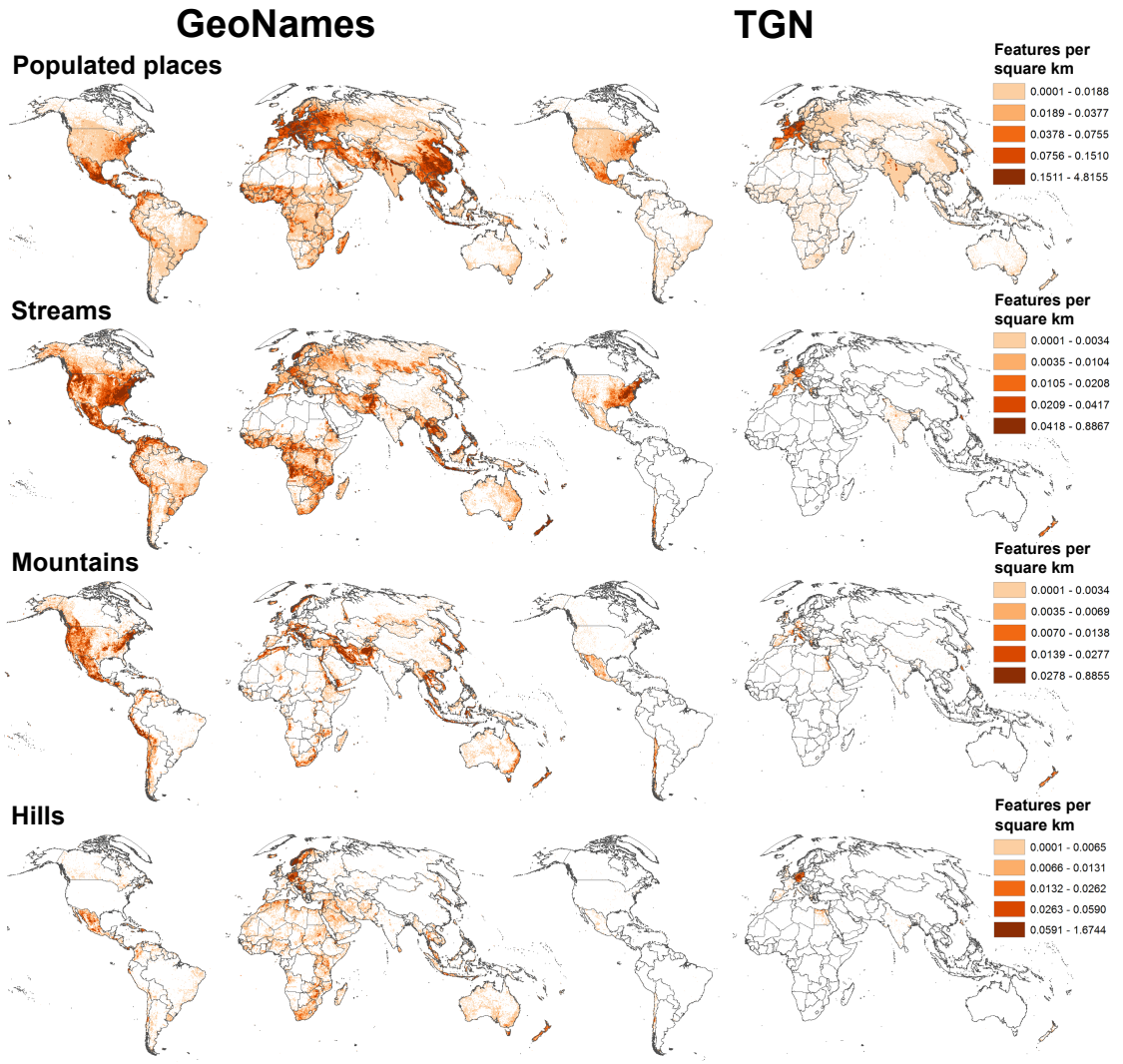


Figure 3.3: Point density maps by gazetteer (GeoNames, TGN) and feature type (populated places, streams, mountains, hills), rendered in terms of GeoNames quantiles, in the Goode Homolosine Land projection. Figure from Acheson et al. (2017a).

the correlation analysis shows strong positive relationships for 'all' and 'populated places' for both the country and raster cell units, and weaker positive relationships for 'mountains'. The relationships for 'streams' are weaker still, with a steep drop in correlation for the country unit as compared to mountains, and the weakest relationship by far (virtually 0) is for 'hills' in the raster unit. Importantly, correlations are greater for countries than raster cells in all five cases examined, showing stronger relationships between record counts in the coarser country unit than for the finer raster cells.

To quantify the relationship between counts in GeoNames and TGN in corre-

Table 3.4: Kendall's tau correlation coefficients between GeoNames and TGN for record counts in corresponding countries and (100x100km) raster cells. Significance levels: * $p < 0.00001$, † $p < 0.01$, $^{+}$ $p < 0.05$. Table adapted from Acheson et al. (2017a).

features	countries (N = 237)		raster (N = 51,996)	
	M (neither 0)	Kendall's tau	M (not both 0)	Kendall's tau
all	237	0.714^{*}	20,665	0.638^{*}
populated places	235	0.701^{*}	14,188	0.538^{*}
streams	29	0.300^{+}	13,182	0.232^{*}
mountains	159	0.485^{*}	9,868	0.245^{*}
hills	74	$0.258^{\dagger}$	8,704	0.042^{*}

sponding cells or countries, we built linear models using GeoNames arbitrarily as the 'independent' variable, and using log-log models because of the positively skewed distribution in both datasets. Our initial scatter plot of the country relationship showed two groups of points, where one group of countries showed higher counts in TGN compared to the others. We called these 15 countries[4] 'high coverage' and introduced a boolean indicator variable in our models representing whether a country was in this set or not. We thus used linear models according to the formula:

$$\ln(TGN) = b_0 + b_1 \ln(GeoNames) + b_2 HighCoverage + \varepsilon \tag{3.1}$$

The final log-log scatter plots for both countries and 100x100km raster cells are shown in Figure 3.4, where we assigned each raster cells to a country (either 'high coverage' or not) or to no country. Our country linear model had an adjusted R^2 of 0.87, hence the record counts in GeoNames could account for 87% of the variation in record counts in TGN. The raster linear model, disregarding cells not in any country, was very similar, with a slightly lower R^2 of 0.82. Record counts in TGN were indeed highly dependent on membership in the 'high coverage' set: for the country model, 'high coverage' countries had around 60 times as many records as low coverage countries ($b_2 = 4.13$, $e^{4.13} = 62.18$) and for the raster model, cells from 'high coverage' countries had around 30 times as many records ($b_2 = 3.40$, $e^{3.40} = 29.96$).

[4]the United States, Germany, Mexico, France, India, Spain, Chile, Taiwan, United Kingdom, Italy, Egypt, Greece, Belgium, New Zealand, and the Netherlands

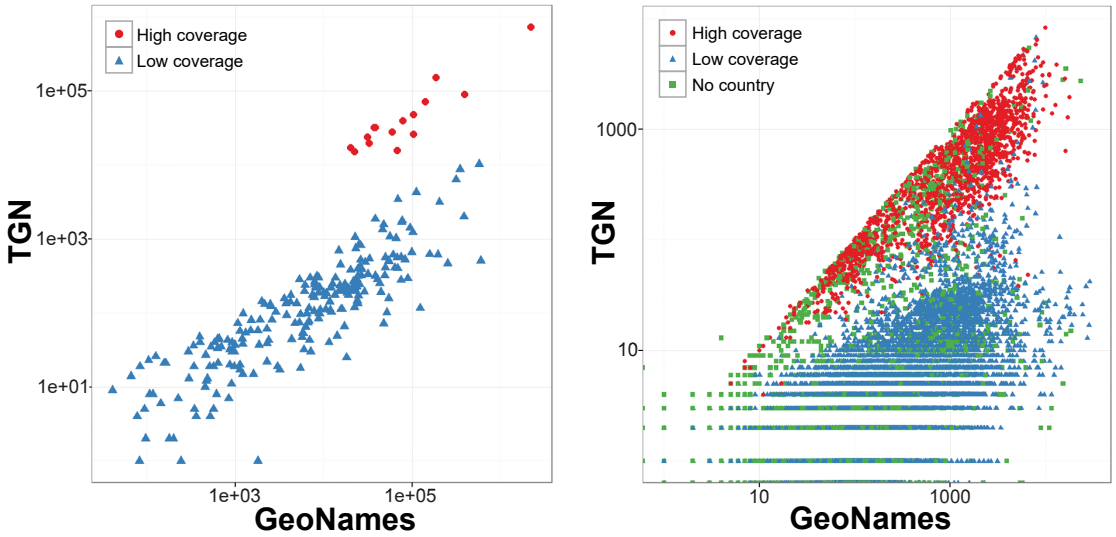


Figure 3.4: Log-log scatter plot of feature counts in TGN as a function of counts in GeoNames in matching countries (left) and 100x100 km cells (right). Figure adapted from Acheson et al. (2017a).

We found that these 'high coverage' countries also had higher record counts than the others in TGN for the four feature type subsets we analyzed. Hence we repeated the correlation analysis (from Table 3.4) but using only the 15 'high coverage' countries. The results, presented in Table 3.5, show that all 5 data subsets have strong positive relationships which are statistically significant. In addition, the correlations are stronger for the 15-country subset than for the full set of countries for the 4 feature type subsets, and the values are quite stable across these different feature types, all ranging from about 0.76 to 0.87. Overall these correlation coefficients show that in the countries where TGN has better coverage, the coverage on the country scale highly resembles the GeoNames data.

Table 3.5: Kendall's tau correlation coefficients between GeoNames and TGN for record counts in countries determined to be in the 'high coverage' group in TGN. Significance levels: * $p < 0.001$. Table adapted from Acheson et al. (2017a).

features	N	Kendall's tau
all	15	0.695*
populated places	15	0.867*
streams	15	0.848*
mountains	15	0.790*
hills	15	0.766*

Summary

Our multi-scale analysis of gazetteer coverage using point density maps, rank correlations, and linear models, shows some consistent patterns: coverage in TGN is much sparser and more idiosyncratic than coverage in GeoNames, coverage of natural feature types is generally more idiosyncratic in both resources than coverage of populated places, and the country unit plays an important role in differences in coverage in both resources.

3.2 Gazetteer matching

After our analysis of coverage in two global gazetteers, we now look at integrating two resources for a defined region of interest and for a subset of natural feature types, which we have seen can have very unbalanced, idiosyncratic coverage in GeoNames and TGN. Specifically, we link or match natural feature records in Switzerland from GeoNames to records from the authoritative swissNAMES3D, when these records are deemed to be about the same real-world entity, a process referred to as gazetteer matching. This type of gazetteer integration can be undertaken before applying a text-to-space pipeline to a particular case study, in order to potentially improve recall at the placename identification or grounding stage through more complete resources, and to improve precision through, for example, accessing records with a richer set of attributes such as alternate names and multiple geometries.

Previous work on gazetteer matching had focused mainly on feature types such as populated places and points of interest (POIs), rather than natural feature types, and strategies were lacking to integrate types into matching decisions when dealing with multiple type hierarchies. More generally, methods and datasets varied from one work to the next, making it hard to evaluate methodological decisions and compare matching performance, and no direct comparisons of rule-based and gazetteer-based methods were found in the literature. In a preliminary work, we annotated and publicly released a dataset matching natural feature records from GeoNames to records in swissNAMES3D, in addition to presenting results of simple

rule-based matching applied to this dataset (Acheson et al., 2017b). The follow-up work summarized here (Acheson et al., 2019) presents a detailed methodology to perform gazetteer matching using machine learning, implements both rule-based methods and machine-learning-based methods for a direct comparison, and highlights ways to construct a realistic testing pipeline, all applied to the publicly released dataset of natural feature records.

3.2.1 Entity resolution and gazetteer matching

Gazetteer matching is a special case of record linking, where the records to be linked represent geographical entities, rather than people, biomedical records, or web pages. Record linking is itself part of the broader challenge of entity resolution (Elmagarmid et al., 2007; Costa et al., 2015; Christen, 2012), which usually consists of a sequence of 3 steps: a data preparation step (cleaning records, normalizing records, aligning fields), a record linking / gazetteer matching step (matching corresponding records), and potentially a record fusion step (building a single resource by merging/augmenting matched records).

Gazetteer matching can itself be divided into 3 steps: candidate selection, feature extraction, and classification. The first step, *candidate selection*, involves choosing a subset of all possible record pairs which will be carried into the next steps. The retained record pairs should ideally include any record pairs that are likely to be matches, so that true matches are found later in the pipeline, but considering *all* record pairs is inefficient and unnecessary since records far apart in space are unlikely to be matches, more or less so depending on the feature type. In the subsequent step, *feature extraction*, various distance (or similarity) metrics (known as 'features' in a machine learning context) are calculated between all candidate record pairs. Examples of matching features include the Levenshtein distance (Levenshtein, 1966) for names and point-to-point distance for point geometries. The final step, *classification*, outputs a 'match' or 'no match' decision for each candidate record pair using a decision boundary set either with rules or based on a

trained machine learning model. An overview of entity resolution, including record linking/gazetteer matching and its steps, is shown in Figure 3.5.

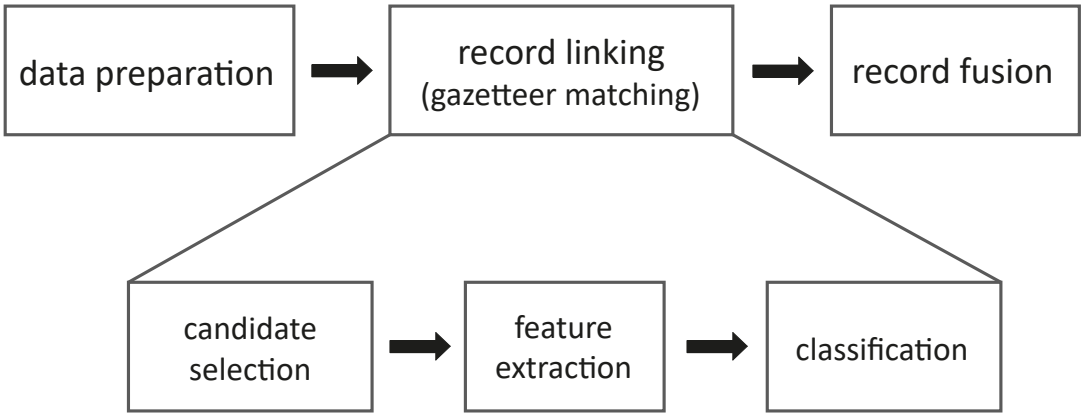


Figure 3.5: Entity resolution steps, including record linking/gazetteer matching, and gazetteer matching sub-steps. Figure from Acheson et al. (2019).

Data heterogeneity commonly results when complex real-world entities get simplified into structured gazetteer records. Records for the same entity in two different gazetteers may have a different name (e.g. a place can have multiple names and variants in different languages), geometry (e.g. two different points or two different geometry types), and type (e.g. types from different types hierarchies, including in different languages). Correspondingly, we focus our matching methods on these three attributes, though we also experiment with additional matching features, particularly ones to help characterize natural features, including elevation and land cover.

3.2.2 Methods

Both rule-based and machine-learning-based matching were implemented, focusing on the core record linking/gazetteer matching step, as laid out in Figure 3.5. Data preparation included projecting coordinates in GeoNames and swissNAMES3D from latitude-longitude coordinates (WGS84) to a Swiss coordinate system (LV03) and vice-versa, in order to more efficiently calculate distances. Record fusion, described above, was not performed for this work. For land cover matching features,

external data was used, consisting of a federally produced 6-class categorization of land cover for the whole of Switzerland[5].

All implementation code was written in Python, relying primarily on the extremely useful *pandas*[6] data analysis library and the *scikit-learn* machine learning library (Pedregosa et al., 2011), and is publicly available[7].

Data

For the matching task, we annotated a subset of records from GeoNames, linking them to records in swissNAMES3D (annotated data made public in Acheson et al., 2017b). We downloaded freely available, daily updated data for Switzerland from GeoNames[8] on 20.07.2017, which contained around 67k records. We also downloaded the freely available swissNAMES3D[9], which is updated annually and fully revised on a 6-year cycle, in February 2017. The dataset contains over 300k records organized according to a Switzerland-specific, German-language feature type hierarchy.

We used GeoNames as our 'source' resource (records to find matches for) and swissNAMES3D as our 'target' resource (to find matching candidate records), since swissNAMES3D is an official dataset and has higher coverage in Switzerland. To manually prepare an annotated gold standard dataset, we first selected a portion of records from our source gazetteer, GeoNames, to find matches for. We randomly selected 50 records per type from each natural feature type which had at least 100 records in Switzerland[10], and comprehensively tried to find matches in swissNAMES3D for these, including one-to-many and 'no match' cases. Each of the 4 annotators, all graduate students in Geographic Information Systems, was assigned all records from two of the 8 feature types to do a first pass annotation of all cases deemed straightforward. Any harder cases were then discussed among

[5]https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/geostat/geodaten-bundesstatistik/boden-nutzung-bedeckung-eignung/arealstatistik-schweiz/bodenbedeckung.html

[6]https://pandas.pydata.org/

[7]https://github.com/eacheson/machine-learning-gazetteer-matching

[8]CH.zip from http://download.geonames.org/export/dump/

[9]from https://shop.swisstopo.admin.ch/en/products/landscape/names3D

[10]The 8 annotated GeoNames types are: lake (LK), glacier (GLCR), stream (STM), peak (PK), pass (PASS), hill (HLL), mountain (MT), and valley (VAL).

the four annotators, until an agreement was reached on each case (as described in Acheson et al. (2017b)). The number of swissNAMES3D matches for each GeoNames record in the annotated data is shown in Table 3.6.

Table 3.6: Number of swissNAMES3D matches for each GeoNames record in the annotated data. Thus, 339 GeoNames records have exactly one match in swissNAMES3D, 14 records have no match, and so on.

number of matches in swissNAMES3D	GeoNames record count
0	14
1	339
2	25
3	13
4	8
6	1

Rule-based matching

We implemented the following set of rule-based matching procedures, from simplest to most complex:

- **random-baseline**: match is a randomly chosen match from all exact name matches.
- **name-threshold**: matches are all exact name matches (primary or alternate names) within a fixed distance threshold (e.g. 5km) of the source record.
- **name-custom-threshold**: matches are as in *name-threshold*, but with type-specific thresholds (c.f. Morana et al., 2014).
- **multi-threshold**: variation on *name-custom-threshold* which adds additional thresholds on land cover and elevation.
- **linear-combination**: combine Levenshtein distance and point-to-point distance (Vincenty, 1975) into an overall score (considering only a subset of records with a compatible type or an exact name match) and keep any matches above a score threshold (c.f. Smart et al., 2010).

We used feature types in two ways in the above rules: to define type-specific distance thresholds in *name-custom-threshold* and *multi-threshold*, and to filter the amount of records we consider in *linear-combination*, the latter being essentially a form of candidate selection. Since combining rules becomes increasingly challenging as their number increases, and simplicity is one of the advantages of rules over machine learning, we manually optimized any thresholds by testing a set of sensible values at fixed intervals, but generally opted for rule sparsity where complexity seemed not to have any evident payoff. An overview of the matching features, as well as a more complete description of the rule-based matching procedures, is available in Paper 2.

Machine-learning-based matching

For machine-learning-based matching, we used random forests (Breiman, 2001) to perform binary classification ('match' or 'no match') on candidate record pairs. Random forests offer many advantages over other potential machine learning algorithms for the gazetteer matching task: they offer high performance with limited overfitting through ensembling, outperforming decision trees and Support Vector Machines on a gazetteer deduplication task (Gonçalves, 2012); they do not require thousands of data points to achieve high performance, as opposed to deep learning approaches; they allow us to work jointly with continuous features (like point-to-point distances) and categorical features (like land cover classes and feature types); and they are fairly simple to use because input features do not have to be normalized and only one key hyperparameter has to be tuned (the number of trees in the forest).

We built a processing pipeline (Figure 3.6) that would approximate a realistic, large-scale matching scenario. This means we split our annotated source records into training and test sets at the beginning of the pipeline in order to handle these two cases slightly differently. In training, we run our candidate selection process, and before running the feature extraction step, we add any known record pairs from our annotated data that our candidate selection process may have missed ('known

matches'), in order to use all the available training data at hand. In testing, we let our candidate selection process choose all the pairs which make it to feature extraction, since in a real matching scenario, there would not be annotated data for new, unseen records. Candidate selection for all records (training and test) was done by applying an initial feature type filter, calculating Levenshtein and point-to-point distance on all remaining record pairs, then keeping the top k pairs on a score combining those metrics (similar to our rule-based *linear-combination* procedure), where k was set experimentally.

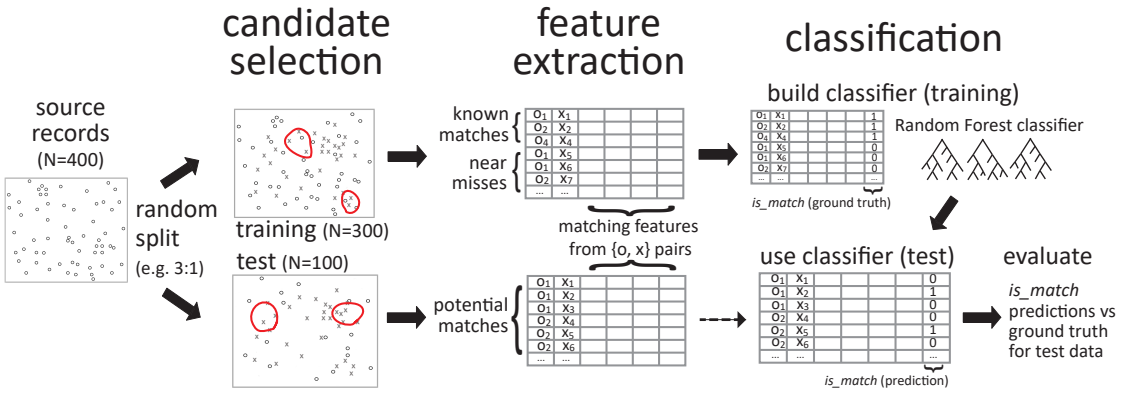


Figure 3.6: Detailed look at the machine learning pipeline for gazetteer matching, including the 3 steps of candidate selection, feature extraction, and classification, with slight differences between the training and testing pipelines. Figure adapted from Acheson et al. (2019).

For the feature extraction step, we used core matching features based on names, geometries, and feature types, with additional features based on elevation and land cover, as with rule-based matching. Since combining either a small or large number of matching features is equally complex in our machine learning pipeline, which is not the case in rule-based matching, we made use of a wider range of name matching features and land cover features. Matching features on names included the Levenshtein distance, the normalized Levenshtein-Damerau distance, the Jaro similarity, the Jaro-Winkler similarity, and the minimum Levenshtein distance considering names and alternate names. Land cover features included the land cover class of the nearest cell, the 'mode' land cover class of the nearest 9 cells, and a feature we call 'land cover distance' which calculates a vector distance between counts of

land cover classes in the source and target records. Categorical information, such as the nearest and mode land cover classes, was turned into a set of matching features using 'one-hot encoding', which converts M categories into M binary features indicating the presence or absence of that category[11]. It is precisely this one-hot encoding that we use to encode feature type information for machine learning, which removes any need to manually align different feature type hierarchies. Instead, each pair is encoded using the same number of binary features, with two of these being non-zero, one indicating the GeoNames type and one the swissNAMES3D type. The random forests can learn statistical associations between the categories/types this way. A more complete, step-by-step description of the machine learning pipeline is available in Paper 2, alongside a more detailed description of the experimental procedures and a full list of the machine learning matching features used.

For the classification step, we tested the following subsets of feature combinations, starting from a *basic* model down to a model using *all* features:

- **basic**: minimum Levenshtein distance and point-to-point distance.
- **str**: all name features and point-to-point distance.
- **basic-type**: minimum Levenshtein distance, point-to-point distance, and one-hot-encoded feature types.
- **str-type**: all name features, point-to-point distance, and one-hot-encoded feature types.
- **str-elev-lc**: all name features, point-to-point distance, elevation, and all land cover features (no feature type information).
- **str-type-lcd**: all name features, point-to-point distance, one-hot-encoded feature types, and land cover distance.
- **all-min**: minimum Levenshtein distance, point-to-point distance, one-hot-encoded feature types, elevation, and land cover distance (uses all attributes but not all features).
- **all**: all features.

[11]For example, with 6 land cover classes, a record in a cell with land cover class 2 could be encoded as [0,1,0,0,0,0], and a record with land cover class 4 as [0,0,0,1,0,0]. Both the category of the source and target record is encoded for each pair.

3.2.3 Results and interpretation

We first present our rule-based matching results, followed by our machine-learning-based results, then briefly probe deeper into matching performance according to feature type and training set size. To evaluate the performance of our rule-based and machine learning based matching, we calculated precision, recall, and $F1$. Precision is the number of positive matches correctly found divided by the total number of positive matches found, while recall is the number of positive matches correctly found divided by the total number of positive matches that were to be found (in the ground truth). There is usually a trade-off between precision and recall, and thus $F1$ combines these two through their harmonic mean and summarizes the overall performance.

Rule-based matching

Our rule-based matching results are presented in Table 3.7. The best performing ruleset is *name-custom-threshold*, with an $F1$ of 0.852, followed closely by the more complex *linear-combination* ruleset at 0.849. The relatively simple *name-threshold* performs quite well, with an $F1$ of 0.830, but the *multi-threshold* ruleset struggles in recall, despite having the highest precision (0.914), which leads to a low $F1$ score overall (0.778). Indeed, simply removing thresholds on land cover and elevation in *multi-threshold*, making it equivalent to *name-custom-threshold*, leads to a higher $F1$ value. Our *random-baseline* ruleset, which simply selects a random match from exact name matches, achieves reasonable precision (0.793), showing that exact name matches play an important role in a subset of our data.

Machine-learning-based matching

Before running our main machine learning experiments, we first fixed a value for k, the number of target records retained for each source record during candidate selection. We did so by testing a range of sensible values over several feature combinations and picking the value ($k = 30$) which lead to the highest $F1$ performances. For our main experiments, we ran the pipeline (Figure 3.6) 20

Table 3.7: Results for rule-based matching, shown for the following thresholds: *name-threshold* distance threshold of 5km, *name-custom-threshold* type-specific thresholds of 5km or 15km (LK, STM, VAL), *multi-threshold* type-specific thresholds of 5km or 15km (LK, STM, VAL), elevation difference threshold of 400m, and land cover distance threshold of 8 units. *random-baseline* results were averaged over 10 runs. Table adapted from Acheson et al. (2019).

name of run	precision	recall	F1
random-baseline	0.793	0.575	0.666
name-threshold	0.876	0.788	0.830
name-custom-threshold	0.843	0.861	0.852
multi-threshold	0.914	0.677	0.778
linear-combination	0.866	0.833	0.849

times, each time with a different 3:1 train-to-test split of the source records. This means a realistic variation of source records are tested and we obtain not only mean values of precision, recall, and F1 per classifier, but also interquartile ranges.

In Figure 3.7, we present the results of these 20 runs, in terms of $F1$, precision, (overall) recall, and classification recall. We calculated two forms of recall: (overall) recall and classification recall. Overall recall represents, given the source records randomly selected for testing at each run, how many positive matches we found out of all of the positive matches in the ground truth. A low recall value here can mean either that we misclassified matching pairs as 'no match', or that our candidate selection was poor and missed many positive matches. Classification recall, on the other hand, measures only classification performance and not candidate selection, by using as a denominator only the matching pairs that made it to the classification stage, rather than the full set of matching pairs in our ground truth. The overall recall is the more meaningful of the two, and indeed the one which factors into $F1$.

In terms of $F1$, Figure 3.7 shows that the 5 right-most feature combinations, which all incorporate feature type information, performed the best (mean and medians between about 0.88 and 0.90), with no clear difference between the combinations. Lower performance is seen for the two simpler combinations with no feature type information (*basic* and *str*), while *str-elev-lc*, which does not have feature type information but does have land cover and elevation information,

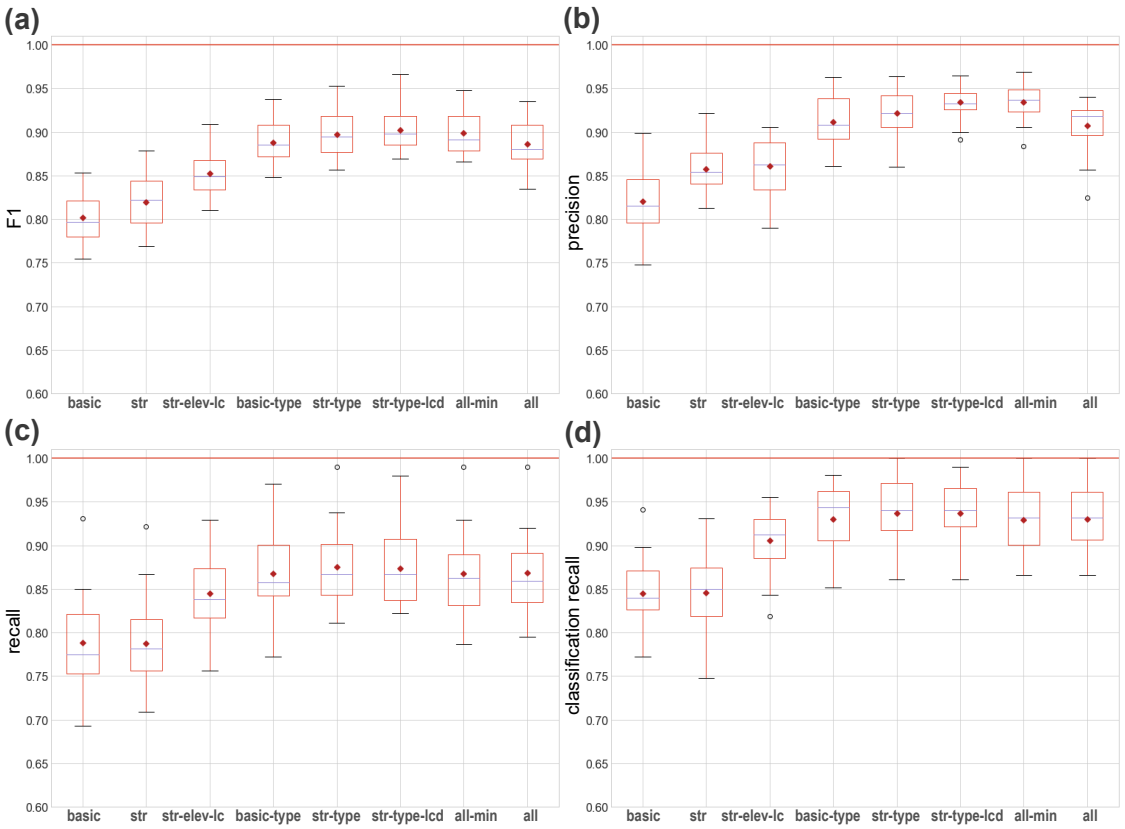


Figure 3.7: Box plot of medians (blue lines) with interquartile range and means (red diamonds) for: (a) F1 (b) precision (c) (overall) recall, and (d) classification recall vs. named combinations of matching features. Figure adapted from Acheson et al. (2019).

performs somewhere in between. The best results over these 20 runs can be observed for the *str-type-lcd*, with the highest $F1$ mean (reaching 0.902), median, upper quartile, and lower quartile.

Further experiments

In order to compare rules and machine learning methods on the exact same test records and break down performance by feature type, we created a fixed, feature-type-balanced test set, consisting of 80 randomly chosen source records, 10 for each of our 8 GeoNames types. We also used this fixed test set to probe further into the performance of our random forests, plotting a learning curve to see how increasing the size of the training data would impact performance. These two sets of results are presented in Figure 3.8.

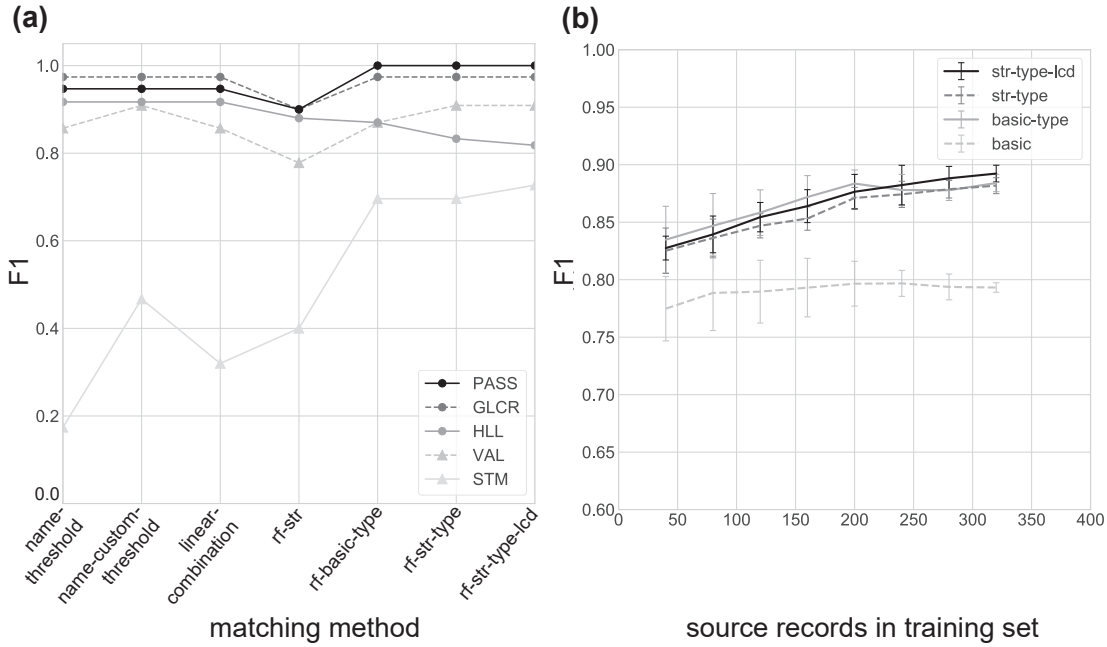


Figure 3.8: F1 performance according to (a) the matching strategy used (from the left, 3 rule-based procedures, in order of increasing complexity, followed by 4 machine learning based methods, prefixed by *rf-*, also in order of increasing complexity) broken down by feature type for 5 selected feature types and (b) the number of source records used in the machine learning training pipeline, showing the mean and standard deviation over 10 runs using incrementally more randomly chosen source records. Figure adapted from Acheson et al. (2019).

Comparing the performance of the 3 plotted rulesets against the 4 plotted machine learning feature combinations (Figure 3.8a), a more balanced performance across feature types can be observed for the machine learning matching, but only when they incorporate feature type information (*basic-type*, *str-type*, *str-type-lcd*). As for the ruleset that incorporates types using type-specific thresholds (*name-custom-threshold*), its performance is actually better on all the plotted types than the machine learning method which does not consider types (*str*). A general takeaway from this plot is that much of the overall performance gain of the better-performing strategies derives from achieving higher performance on the most 'difficult' type (STM), without compromising performance on other types.

The learning curve (Figure 3.8b) shows that as we increase the size of the training set (by increments of 40 source records, up to the maximum of 320 (400-80)), the $F1$ performance continually increases for the feature combinations *str-type-lcd* and

str-type. The combination *basic-type* increases up to 200 source records, then appears to plateau, though maintaining strong performance, while the *basic* combination, which only considers names and geometries, offers lower performance and also plateaus around 200. The learning curve suggests that additional performance gains could be achieved (with *str-type-lcd* and *str-type*), were there additional data to use to train the random forests.

Summary

Gazetteer matching is an important task central to creating optimal placename resources customized for particular text-to-space applications. To match natural feature records from GeoNames to swissNAMES3D, we implemented and compared both rule-based and machine-learning-based matching, the latter using random forests, which were particularly well suited to the task, offering high performance with (only) hundreds of annotated pairs, joint handling of continuous and categorical features, and low levels of required pre-processing and tuning. Rule-based methods showed satisfactory performance, especially the *name-custom-threshold* ruleset (considering record names, geometries, and types via type-specific distance thresholds), which was both simple to implement and gave the highest $F1$ performance on our data (0.852). Our best machine learning models, however, offered a 6% increase in $F1$ performance over this best ruleset, achieving a mean $F1$ of 0.902 in the case of the *str-type-lcd* feature combination, which incorporated matching features on names, geometries, feature types, and land cover.

Generally, for our natural feature records, incorporating feature type information into matching decisions was crucial. Indeed, rule-based and machine learning performance was similar when considering only record names and locations, but all machine learning models incorporating feature type information achieved better mean $F1$ values (> 0.88). One-hot encoding of types enabled random forests to consider type alignment in a data-driven fashion for classification, without having to manually align type hierarchies, whereas incorporating types into rulesets was more complex, requiring some form of manual type alignment and generally decisions

tailored to a particular dataset. Machine learning matching thus offered higher performance than rules, at the cost of greater - but stable - complexity. Rules on the other hand were simpler to implement, but with complexity increasing with the number of matching attributes to consider. Rules would additionally be harder to generalize to a new dataset, as several thresholds were manually set based on our particular data. In a new matching scenario, the performance edge of machine learning over a well-crafted ruleset would depend on the specifics of the datasets to match, but given the very large number of records often at stake, this initial extra implementation effort should in most cases pay off.

The most obvious characteristic of science is its application: the fact that, as a consequence of science, one has a power to do things.

— Richard P. Feynman

4

Text-to-space applications: case studies

Contents

The contents of this chapter are based on two publications included in this thesis: Paper III (Wartmann et al., 2018) and Paper IV (Acheson and Purves, submitted).

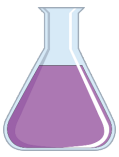
This chapter concerns the second theme of this thesis: text-to-space *applications*. Studying text-to-space pipelines is most insightful in an applied context. Indeed, processing decisions are influenced by the particularities of the corpora, resources, downstream tasks, and, as we discussed in the context of gazetteer matching, reward-to-effort considerations. Minimally, building a text-to-space pipeline involves deciding on a strategy to identify locations in text, a strategy to disambiguate and ground these locations using suitable (potentially integrated) resources, and a way to represent the documents that serves the needs of the application.

Hence, in the context of this thesis, text-to-space pipelines were built for

practical purposes in two contrasting case studies. The first case study (section 4.1) generates areal 'footprints' from a German-language, semi-formal corpus of hiking blogs gathered for specific study sites in Switzerland. Within the broader goal of studying how people perceive and characterize landscapes through language, three complementary textual sources are analyzed and compared (free lists, hiking blogs, and Flickr tags), and the text-to-space pipeline enables the spatial querying of data based on data-driven footprints representing each study site. The second case study (section 4.2) extracts and represents locations from scientific articles, where one of the main challenges is to identify relevant locations, such as study sites or treatment locations, while filtering irrelevant locations, such as company headquarters or locations in references. A fully automatic text-to-space pipeline is built which starts from PDF files and outputs structured location information alongside corpus maps. The pipeline is applied to, and minimally customized for, two scientific article corpora, exploring how much tailoring is needed for the domain (ecological vs biomedical domain), within a particular text genre (scientific articles).

This chapter summarizes the methods and results from these two contrasting case studies. An overview of the characteristics of the case studies is presented in Table 4.1.

Table 4.1: Overview of the two case studies.



	hiking blogs	**scientific articles**
corpus size	50	350
language	German	English
style	semi-formal	formal (science)
corpus coverage	Switzerland	global

4.1 Case study I: hiking blogs

Our first case study builds a processing pipeline to generate spatial footprints for a spatially-rich, fine-granularity corpus of hiking blogs. The pipeline is part of a larger methodological project on how to capture rich landscape descriptions from people in Switzerland, with the goal of going beyond the many studies using one type of information source (such as social media or in-person interviews). Instead, in the main study summarized here (Wartmann et al., 2018), we characterize and compare landscape descriptions from three different sources: free lists from in-person interviews, social media data in the form of Flickr photo tags, and hiking blogs crawled from the web. We generate geographically-focused areal footprints from the hiking blogs, one footprint for each of our ten study sites, and we use these footprints to query Flickr photos based on location. By using textual data sources that are linked in geographical space, we can then examine how the geographical area, landscape type, and the data source itself relate to the information content in our data. We obtain focused study site footprints by filtering outliers from all the grounded placenames for a particular study site, and in a follow-up work (Acheson et al., 2017c), we refine this footprint generation process to include point clustering using DBSCAN (Ester et al., 1996) (short for Density-Based Spatial Clustering of Applications with Noise), an algorithm designed to efficiently find clusters of arbitrary shape with only few input parameters.

4.1.1 Methods

We first give a brief overview of the methods of the landscape study which motivated the development of a text-to-space pipeline, then we explain the processing decisions made while building the pipeline.

Landscape study

The broader project on studying landscape characterizations in textual sources focused on ten specific study sites in German-speaking Switzerland. These sites were selected at the start of the project based on their popularity, their landscape

type (based on a formal landscape typography for Switzerland (ARE, 2011)), and their geographical location (to get a reasonable spread of sites which are accessible through public transport and hiking trails). An overview of the study sites is shown in Figure 4.1. The project leader, a native speaker of Swiss-German, interviewed 30 people at each of the ten pre-determined locations, having each respondent perform a 'free listing' exercise (described further in Wartmann and Purves, 2018) to elicit terms associated with the landscape. The responses were transcribed, in order, and this set of *free lists* formed the first of three data sources that we analyzed.

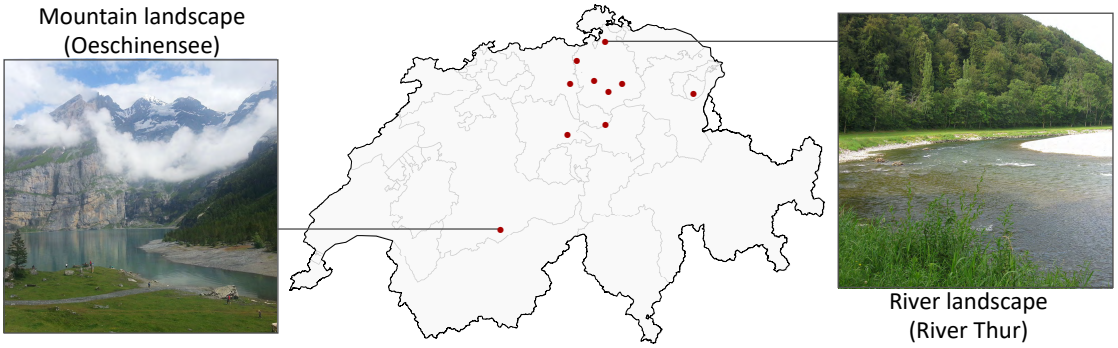


Figure 4.1: Overview of the ten study sites, showing two example sites: a mountain landscape (Oeschinensee) and a river landscape (River Thur). Photo credits: Flurina Wartmann.

The second data source was obtained through targeted web crawling, using the BootCaT platform (Baroni and Bernardini, 2004). Using query terms consisting of study site toponyms alongside 'wandern' (hiking) and 'wir' (we), sets of documents were returned and triaged down to 5 first-person narratives per study site. These 50 documents formed our second data source for the landscape study, *hiking blogs*, as well as the input to our text-to-space pipeline. Indeed, in order to obtain relevant data for our third data source, *Flickr tags*, some form of areal footprint for each study site was required to then use to spatially query Flickr data. We thus built a text-to-space pipeline to generate these areal footprints so that each footprint would reflect the first-person descriptions we had collected and consider the landscape setting of each site. This was deemed a superior option to drawing arbitrary boundaries ourselves or, for example, defining a fixed radius around the interview locations.

With three types of text documents and ten study sites, we could then compare study sites, landscape types, and data sources using cosine similarity between documents represented as term vectors (Manning and Schütze, 1999). Furthermore, by using a coding scheme to classify terms into categories (or aspects) such as 'toponym', 'sense of place', and 'biophysical', we could tailor our comparisons to particular aspects. For example, we could ask whether different landscape types lead to significantly different text descriptions with respect to 'sense of place' aspects or 'biophysical' aspects, by building term vectors from just the terms classified as the aspect in question. We assessed whether comparisons of groups of cosine similarity values were statistically significant using two-sided Mann-Whitney-U tests (significance level $\alpha = 0.05$). An overview of the methodological sequence of the landscape study is shown in Figure 4.2. Next, we take a detailed look at our text-to-space-pipeline.

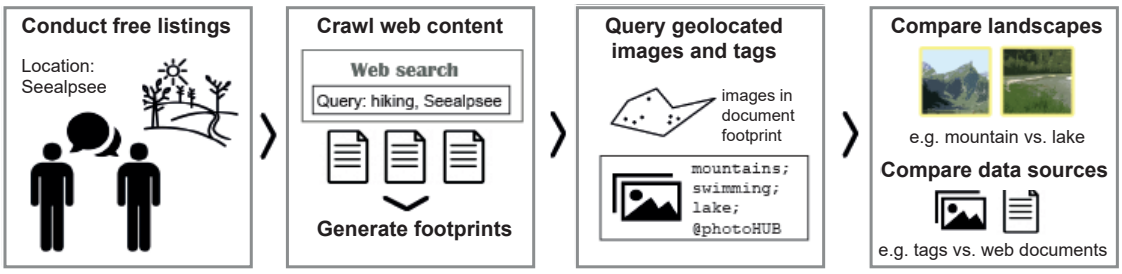


Figure 4.2: Methodological overview of the landscape study, showing the context for footprint generation from web-crawled hiking blogs. Figure from Wartmann et al. (2018).

Text-to-space pipeline

Our pipeline followed the 3-step model described early in this thesis: 1. identifying placenames, 2. grounding placenames, and 3. building document-level geographic representations. An overview of these steps applied to the current case study is presented in Figure 4.3. For the first step, placename identification, we decided to identify placenames in the hiking blogs manually. Several reasons motivated this decision: our corpus was small enough that this was feasible and possibly less time-consuming than implementing an automatic solution; we estimated that high-performance automatic placename identification would be challenging for our

German-language semi-formal blogs, which included non-standard spellings and featured many fine granularity placenames; poor placename identification has been reported as a major source of error in text-to-space pipelines which can propagate downstream (Amitay et al., 2004; Purves et al., 2007); and our particular task required as many (correctly identified) placenames as possible in order to form detailed polygonal footprints for the downstream query task.

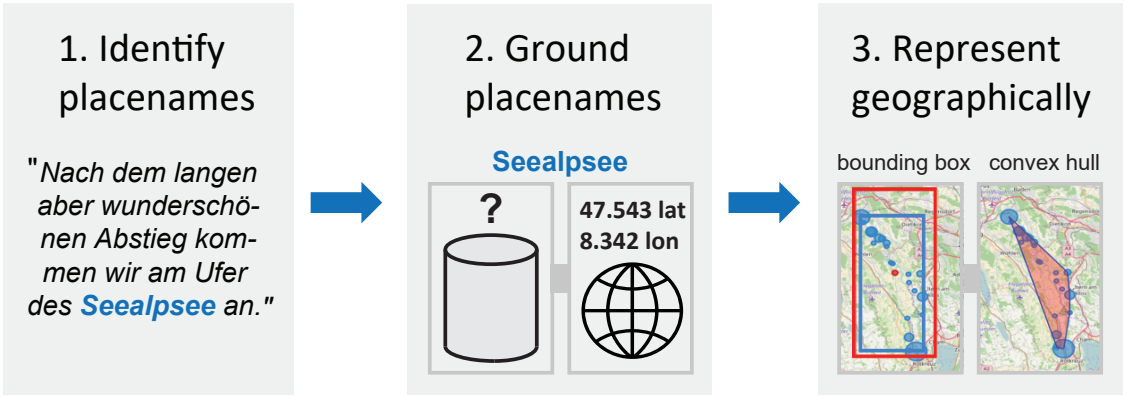


Figure 4.3: Three-step text-to-space pipeline applied to generate footprints from hiking blogs.

For the second step, grounding placenames, we relied on the GeoAdmin API[1] to obtain ranked results from swissNAMES3D for each identified placename. Using swissNAMES3D as our gazetteer resource was the obvious choice for our Swiss hiking blogs, given its high quality, easy access, and detailed coverage of Switzerland including natural features. Accessing the resource using the well-developed GeoAdmin API allowed us to immediately circumvent some complexities in the raw gazetteer data, such as dealing with names joined with cantonal abbreviations (e.g. 'Pfäffikon, SZ') and having to create our own result ranking functions. We used the API's 'location search' feature to query each placename, after compiling a list of the toponyms and their counts for each study site (i.e. within 5 hiking blogs), which returned structured information including point coordinates for each result. We performed no explicit disambiguation at this stage, instead keeping the top result for each toponym.

[1]https://api3.geo.admin.ch/ (accessed in 07.2019)

For the final step, geographically representing the documents, we implemented two approaches to generate the final set of points for footprints: iterative filtering of outliers based on the centroid and standard deviation of our candidate points (Smith and Crane, 2001), and clustering using DBSCAN to identify one main cluster and thereby discard outliers (Moncla et al., 2014b). In the filtering approach, the mean and standard deviation of all the remaining points (one per placename) are calculated in a loop, and these dispersion metrics are used to discard points (placenames) which are further away than most, eventually homing in on the study site. Several parameters were heuristically set for the filtering: a minimum and maximum number of iterations to run, a maximum size for the height or width of the bounding box, and a scaling value for the standard deviation. In the clustering approach, DBSCAN was used to identify a main cluster of points, considering the scale of our study sites and hiking blogs. Two parameters were set for DBSCAN: ϵ (the maximum distance for points to be in the same cluster), and *minPts* (the minimum number of points in a cluster). Once points were either filtered or clustered, the points remaining or within the main cluster were used to construct bounding boxes and convex hulls for each study site[2].

The entire processing pipeline was automated in Python, starting from the manually extracted list of placenames for each study site. Python libraries used include the *arcgis* library[3] to generate convex hulls, the *geopandas* library[4] to work with spatial data in tables, the *scikit-learn* library for DBSCAN, and the *folium* library[5] to generate maps.

4.1.2 Results and interpretation

We first report on the results of our text-to-space pipeline, then briefly cover the main results from the broader landscape study.

[2]The filtering-generated convex hulls were used in the landscape study for timing reasons.
[3]https://developers.arcgis.com/python/ (accessed in 07.2019)
[4]http://geopandas.org/ (accessed in 07.2019)
[5]https://python-visualization.github.io/folium/ (accessed in 07.2019)

Text-to-space pipeline

With our manually annotated placenames and high quality placename resources (GeoAdmin and swissNAMES3D), simple approaches to placename grounding worked well enough for our purposes: with the numerous placenames identified for each study site, we had a high enough 'signal' to build focused footprints by retaining just the top candidate for each placename. We experimented with slight changes in our processing chain including: how many candidates to retain per placename; whether to give higher priority to placenames with exactly one candidate (and hence, 'unambiguous' with respect to our gazetteer resource); and whether to give more weight to placenames appearing multiple times within a study site. None of these changes yielded obvious gains, hence we stuck to the simpler approach. Importantly, the placename density around study sites was high enough that our filtering and clustering approaches could succeed.

For the geographic representation step, we obtained satisfactory results using both filtering and DBSCAN clustering, but the DBSCAN results were better suited to more complex geometric arrangements and required less parameter tuning. Indeed, the filtering approach required several parameters to be experimentally set in order to obtain a suitable set of footprints, including escape conditions on the loop (to run between 3 and 10 iterations and to stop when we went below a bounding box dimension threshold) and scaling the standard deviation (by a factor of 2). The clustering approach, on the other hand, had built-in escape conditions and required setting just two parameters for DBSCAN: the maximum distance for points to be in the same cluster (5km), and the minimum number of points in a cluster (3). Figure 4.4 shows the application of DBSCAN clustering to all 10 study sites.

Landscape study

In the broader landscape study, we looked at how the textual data we gathered about landscapes related to the study sites (10), data source types (3), and landscape types (5). We found that documents of the same *data source* were significantly more similar than documents from different data sources, irrespective of the study

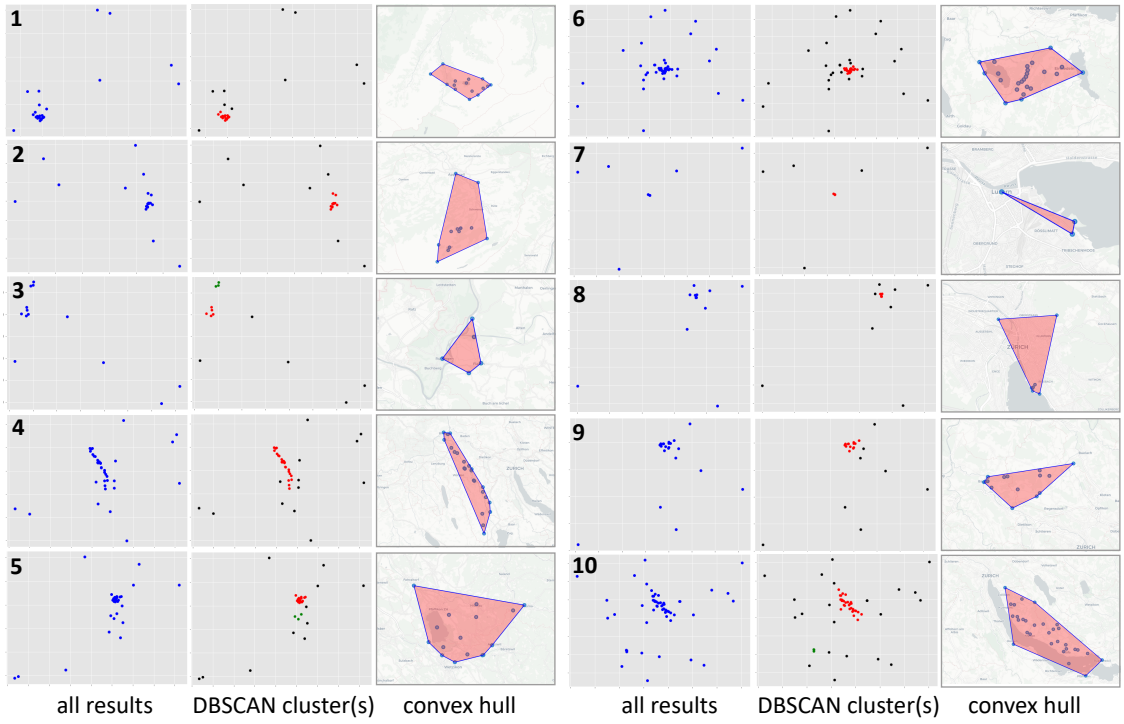


Figure 4.4: Example of footprints obtained using DBSCAN clustering, for each study site, showing the input points to DBSCAN (all top results from geocoding), the cluster(s) returned by DBSCAN (main cluster in red), and the convex hull around the main cluster.

site or landscape type they were located in. Thus, the data source itself has a big influence on the way that landscapes are characterized in text. We then looked at whether landscape descriptions were more similar when they were from the same *landscape type*. Here, most comparisons were not significant. However, when we controlled for the influence of the data source, by comparing documents of a particular data source within and across landscape types, we obtained significant comparisons for two term subsets: landscape terms with toponyms removed, and terms classified as biophysical aspects. For terms classified as 'sense of place', these comparisons were not significant. Hence, in our study, Swiss-German people described landscapes differently in terms of their biophysical properties (such as cliffs and crevasses in a mountain landscape vs. flowing water and woods in a river landscape), which relates to the landscape type. However, they described a similar sense of place even in different types of landscapes, which suggests that natural landscapes which are visited for recreational purposes may evoke similar feelings

of identity, relaxation, and a connection to nature.

Summary

The downstream task of spatially querying for Flickr data largely set the requirements for our footprint generation pipeline. An important factor for the success of our pipeline was the availability of quality placename resources for our area of study. Our decision to manually annotate placenames was influenced by a desire to prioritize getting results for our study over conducting methodological research on placename identification. However, it would of course be interesting to automate this placename detection process and, within the constraints of the corpus properties (German, semi-formal text), it would make the text-to-space pipeline more scalable and applicable to new cases. The purpose of our study was to gather and compare landscape descriptions, and, ultimately, a true measure of success of our text-to-space pipeline is the success of the landscape study itself: we were able to test our hypotheses, obtain intuitive results, and provide new insights for future work on landscapes and text.

4.2 Case study II: scientific articles

In our second case study, we build a fully automatic processing pipeline to extract and represent relevant locations from two corpora of scientific articles. In contrast to the first case study, a major focus of this work is the first step of the 3-step text-to-space pipeline: identifying locations. Indeed, automatically identifying locations from these articles involves several sub-steps, including converting our input PDF files to unstructured and semi-structured text formats, cleaning this text, extracting text portions likely to contain locations of interest, running an NER tool over these text portions, and finally dealing with the output of the NER tool in a way that maximizes performance on our task (that is, identifies correct, and ignores incorrect, locations). A key part of this location identification step is to filter the location candidates we identify, such that we retain relevant content locations such as study sites and patient treatment locations, and discard irrelevant

locations such as locations from cited studies, locations in references, and locations indicating where a company providing commercial products is based.

Work on extracting geographical locations from scientific articles has so far been relatively rare. Several works have focused on the problem of extracting meaningful locations such as study sites from scientific article collections, usually within the ecological domain, but most of these have used manual approaches to identify locations (Wallis et al., 2011; Martin et al., 2012; Karl et al., 2013; Margulies et al., 2016), with a few works using semi-automatic (Fisher et al., 2011) and automatic approaches (Tamames and de Lorenzo, 2010; Leveling, 2015; Kmoch et al., 2018). A related stream of work that has recently gained traction as a SemEval-2019 task (Weissenbacher et al., 2019) aims to perform comprehensive toponym recognition and resolution in scientific articles, specifically on an annotated corpus in the domain of phylogeography (Weissenbacher et al., 2015; Tahsin et al., 2016; Weissenbacher et al., 2017; Magge et al., 2018). However, these works identify *all* toponym mentions within the main text of an article, rather than a subset of relevant locations, and hence an important part of our own research problem is missing. Indeed, our task is more closely related to automatizing what a human annotator would extract for a meta-analysis considering geography, and we aim to create results which could be immediately useful and/or efficiently reviewable by a human annotator. Below, we present our two corpora, the details of our text-to-space pipeline, and our results including maps of the extracted locations.

4.2.1 Corpora

We developed and tested our pipeline on two corpora of articles from different scientific domains: one from the ecological domain and one from the biomedical domain. The ecological corpus, *Orchards*, was an early, minimally triaged collection of articles relating to fruit orchards. The articles were collected for a meta-analysis on agricultural practices and biodiversity, which focused on Mediterranean climates (van der Meer et al., 2017). The biomedical corpus, *Cancer*, was a

random subset of articles from the database Progenetix[6] which had a PDF file available. Progenetix is a curated cancer genomics database focused on compiling information on Comparative Genomic Hybridization and Whole Genome/Exome Sequencing studies (Cai et al., 2014). We manually annotated 150 Orchards articles and 200 Cancer articles, and in both cases set aside 50 articles for evaluation, while using the remaining articles to iteratively develop our pipeline. Summary information is presented in Table 4.2.

Table 4.2: Summary information about the two corpora.

corpus		articles (annotated)		
name	domain	total	train	test
Orchards	ecology	150	100	50
Cancer	biomedical	200	150	50

For each article, the information we annotated included: the ground truth study/sample/patient locations (if any), where this information was found in the article, the quality of the location information (the annotation categories are described in the Figure 4.5 caption), and the publication year of the article. The articles in the Cancer corpus often showed missing or poor location information, whereas the location reporting was consistently good in the more spatial Orchards corpus (Figure 4.5 (a)). With respect to publication years, it is the Orchards corpus that showed greater variance. Indeed, the Orchards corpus contained some articles published in the 1970s and 1980s, which could complicate text extraction, whereas the oldest article in the Cancer corpus was from 1995, consistent with the corpus' focus on scientific techniques which were developed in the 1990s (Figure 4.5 (b)).

4.2.2 Methods

We developed a fully automatic text-to-space pipeline which extracts and represents relevant locations from scientific articles using freely available tools combined with rule-based processing. It takes PDF files as input and outputs structured location

[6]https://progenetix.org/

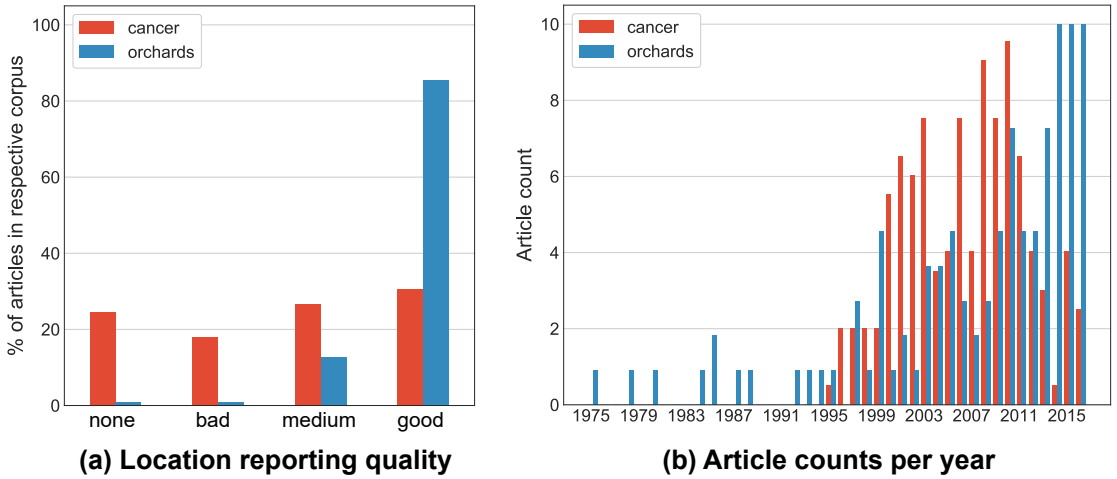


Figure 4.5: Comparison of corpora according to (a) location reporting quality and (b) publication year. Categories for location reporting quality (a): none: no mention of study/sample location; bad: implicit location info (such as 'our institute') or reference to another paper; medium: study/sample location info like name of institute only and perhaps some locations not mentioned (incomplete location info); good: explicit study/sample location info that could probably be extracted and geocoded (such as mentioning a city and country). Figure from Acheson and Purves (submitted).

information for each file, including a location string, its sentence context, and point coordinates for each location. An overview of the pipeline is shown in Figure 4.6.

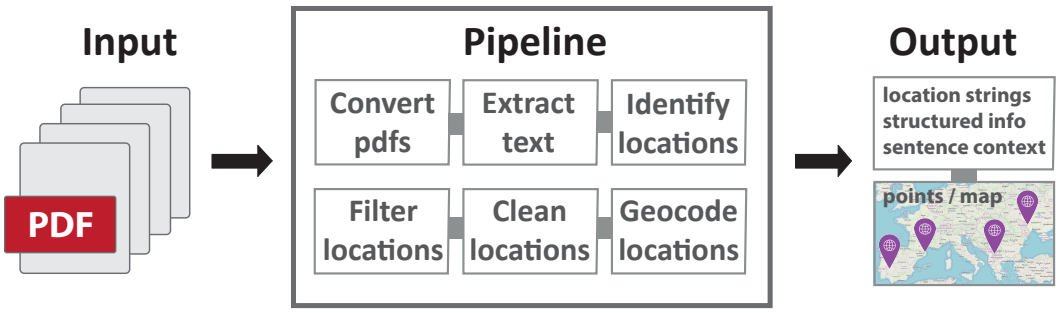


Figure 4.6: Overview of the automatic processing pipeline. Figure from Acheson and Purves (submitted).

We developed the pipeline iteratively on our training articles and aimed to minimize domain-customization across the two corpora. We limited extracted locations to relevant (study/sample) locations in two main ways: 1. by only looking for locations in targeted portions of the article (pre-NER processing) and 2. by filtering identified locations to exclude company locations and other irrelevant locations (post-NER processing). Indeed, the presence of irrelevant locations

throughout scientific articles is cited as a major obstacle to automatically extracting placenames in Karl (2018), who instead focus on extracting coordinates, and has led to poor performance on full-text articles (Kmoch et al., 2018). The core of our location identification strategy relies on using an out-of-the-box NER tool, Stanford NER, which has outperformed the competition on recent multi-dataset comparisons[7] (Jiang et al., 2016). The output from this NER tool is then chunked and filtered before the location grounding step. Our location grounding strategy relies on another high-performing tool, the Google Geocoding API[8], to get structured information and a spatial representation for our extracted location strings. The geocoding results include a fully qualified location string, a return type with granularity indications, and a latitude and longitude which can easily be mapped.

The detailed steps of our processing pipeline (Figure 4.7) are as follows:

- **Convert PDFs:** Each PDF document is converted to 1. a plain text file, using pdfminer[9], and to 2. an XML file using CERMINE (Tkaczyk et al., 2015), a Java-based library to extract metadata and contents from scientific article PDFs. Performing both file conversions provides the pipeline with redundancy in the case of failed or error-rich conversions.
- **Extract text:** Portions of the article contents likely to contain relevant locations are extracted by identifying relevant headings (such as methods or study site sections) using regular expression matching.
- **Identify locations:** The extracted text portions are pre-processed (cleaned, tokenized, and POS-tagged), then NER-tagged using Stanford NER (3-class classifier), accessed from the NLTK Python library (Bird et al., 2009) (Stanford NER v3.8.0, NLTK v3.2.5). The NER output is then chunked using custom code to identify location units, aiming for high recall (missing as few true locations as possible).

[7]It also outperformed spaCy in our own pilot testing on a subset of training articles.
[8]It outperformed OSM Nominatim in our pilot testing on a subset of extracted location strings.
[9]https://github.com/pdfminer/pdfminer.six

- **Filter locations:** The location units identified in the previous step are filtered using rules to remove any candidate that is not deemed a relevant location, including non-locations, suspected company locations, and citations. Rules consider: tag sequences (e.g. reject candidates with no 'location' tags), presence of keep words (e.g. keep candidates with 'University' or 'Institute'), presence of discard words (e.g. reject candidates with 'Inc' or 'GmbH'), and token (tag, word) combinations. The goal of this step is to increase precision, while trying to maintain good recall. This step produces our final list of identified content locations.
- **Clean locations:** Content location strings are cleaned of any trailing prepositions or punctuation before geocoding.
- **Geocode locations:** Clean location strings are sent to the Google Geocoding API and the top result is retained for each.

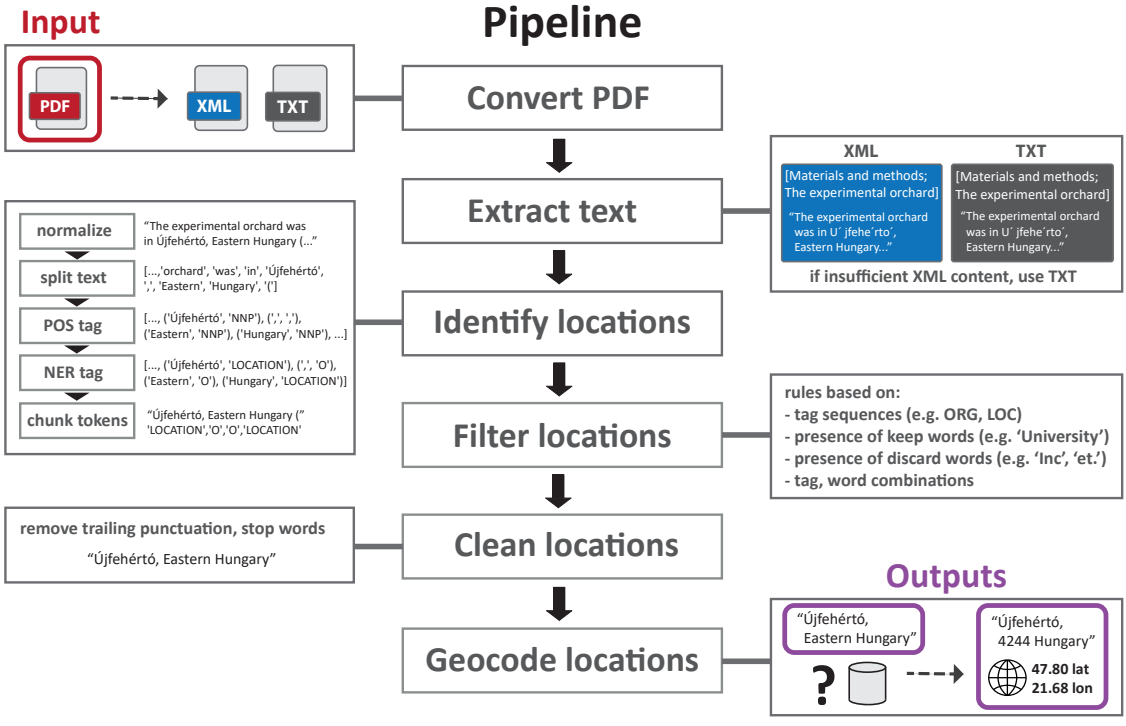


Figure 4.7: Detailed look at the processing pipeline. Figure from Acheson and Purves (submitted).

For domain-adaptation, we minimally customized the regular expressions to detect relevant section headings (*Extract text* step) and the location chunking

(*Identify locations* step). Indeed, relevant headings in the Orchards corpus featured words like 'region', 'area', and 'site', compared to words like 'patient', 'sample', 'specimen', and 'subject' in the Cancer corpus. As for location chunking, location strings in the Orchards corpus often contained cardinal direction words (like 'east' and 'southern') and geographic entity type words (like 'region', 'county', and 'park') and we decided to include these words in our final location strings in order to keep location 'units' together (such as 'Nancy (East of France)' instead of 'Nancy' and 'France'). We found this tended to give better context for the geocoding step and had an overall positive effect on performance.

Finally, to show our results, we programmatically generated interactive maps of the locations found in the test articles by mapping the coordinates obtained at the *Geocode locations* step, when a result was returned. As in the hiking blogs case study, the *folium* library was used to generate maps in Python.

4.2.3 Results and interpretation

The results from running our processing pipeline on the two sets of test articles[10] are presented in Table 4.3 (detailed explanations follow). Overall, results were slightly better on the Orchards corpus than on the Cancer corpus, with the exception of the geocoding accuracy which was higher for the Cancer corpus. Higher location extraction performance was somewhat expected given the superior location reporting quality in the Orchards corpus (Figure 4.5 (a)).

The outputs from our pipeline were evaluated both in a location-centric way (*location unit*), to isolate different parts of our pipeline, and in an article-centric way (*article unit*), to get a balanced picture of performance. In the location-centric evaluation, we separately evaluated our two main outputs, location strings (*extraction precision*) and geocode results (*geocode accuracy*), as well as their sequence (*full pipeline precision*). Geocode accuracy was calculated on the subset of

[10]For the Orchards corpus, the results are for a subset of articles which were studies, rather than review articles, editorials, or articles in popular science magazines. These studies formed between 73-74% of articles in the full minimally-triaged collection, training set, and test set. Results from the full corpus were very similar and are available in the supplementary materials of the submitted article.

Table 4.3: Results of our processing pipeline, aggregated either with respect to extracted locations (*location unit*) or articles (*article unit*). Table from Acheson and Purves (submitted).

	location unit			**article unit**		
	extraction	geocoding	full pipeline	extraction (weighted)		
corpus	precision	accuracy	precision	precision	recall	F1
Orchards	0.869	0.906	0.842	0.822	0.804	0.813
Cancer	0.810	0.980	0.778	0.740	0.769	0.754

true positive extracted location strings, whereas full pipeline precision was calculated for all extracted location strings. In the article-centric evaluation, we calculated both extraction precision and recall out of a maximum value of 1 for each article, in order to not give a disproportionate amount of weight to articles with multiple extracted location strings (precision) or multiple study sites (recall). We then summed these per-article values and divided them by the number of articles to get final precision, recall, and $F1$ (harmonic mean of precision and recall) values for the *article unit* (Table 4.3).

We also conducted a detailed error analysis on the test corpora, classifying all errors into categories, with counts presented in Table 4.4 and examples of each category in Table 4.5. The most frequent errors were NER errors, then text extraction errors, geocoding errors, and comma group errors[11]. Geocode errors were almost all in the Orchards corpus, possibly because locations in the Orchards corpus tended to be more ambiguous and less 'important' globally (smaller, less populated locations) than in the Cancer corpus (where genome work is largely done in major hospitals and research institutes in big cities). Though location strings not disambiguated by a city or country appeared in both corpora, the Google Geocoding API still mostly gave correct answers for unqualified strings in the Cancer corpus (e.g. 'Massachusetts General Hospital', 'Royal Free Hospital and Medical School') but not in the Orchards corpus (e.g. 'Via Emilia', 'Dry Creek Vineyard').

[11] A comma group error occurred when multiple locations were separated by commas and were falsely chunked as a single location unit by our code (e.g. 'Burlington, Cambridge', instead of 'Burlington' and 'Cambridge').

Table 4.4: Errors in both corpora classified into categories, shown as raw counts and as the percentage of the total errors for that corpus. Table from Acheson and Purves (submitted).

	Orchards		**Cancer**	
error description	count	percent	count	percent
NER error	12	27.3	8	32.0
text portion not extracted	8	18.2	7	28.0
wrong/no geocode result	9	20.5	1	4.0
comma group	7	15.9	0	0.0
candidate filtering error	3	6.8	4	16.0
non-standard headings	3	6.8	0	0.0
other	2	4.5	5	20.0
total	44	100	25	100

Table 4.5: Errors examples for each error category. Table adapted from Acheson and Purves (submitted).

error description	example
NER error	Rome as location in 'MacIntosh or Rome varieties'
text portion not extracted	location only appears in Acknowledgements
wrong/no geocode result	'Moldova Region' in Romania not Moldova
comma group	'Burlington, Cambridge' taken as one location
candidate filtering error	company location not filtered out
non-standard headings	'Almonds' sub-heading contained study site info
other	wrongly extracted publisher location in footer

Finally, screenshots of the interactive global maps of geocoded locations in the two corpora are shown in Figure 4.8. Since all geocode results were mapped, including ones for wrongly extracted location strings[12], the color-coding represents the full pipeline precision from Table 4.3. Results for a new, unevaluated corpus could be mapped using a single color, providing a visual estimate of coverage prior to evaluation/validation. Both maps are dominated by locations in Europe and North America, showing where most studies or samples were conducted or collected, but the false positives (red), predominately in North America, likely reflect an underlying tendency of the geocoder to default to locations in North America, which may reflect the underlying gazetteer data (c.f. Acheson et al., 2017a).

[12]Note that not all wrongly extracted location strings had a geocode result.

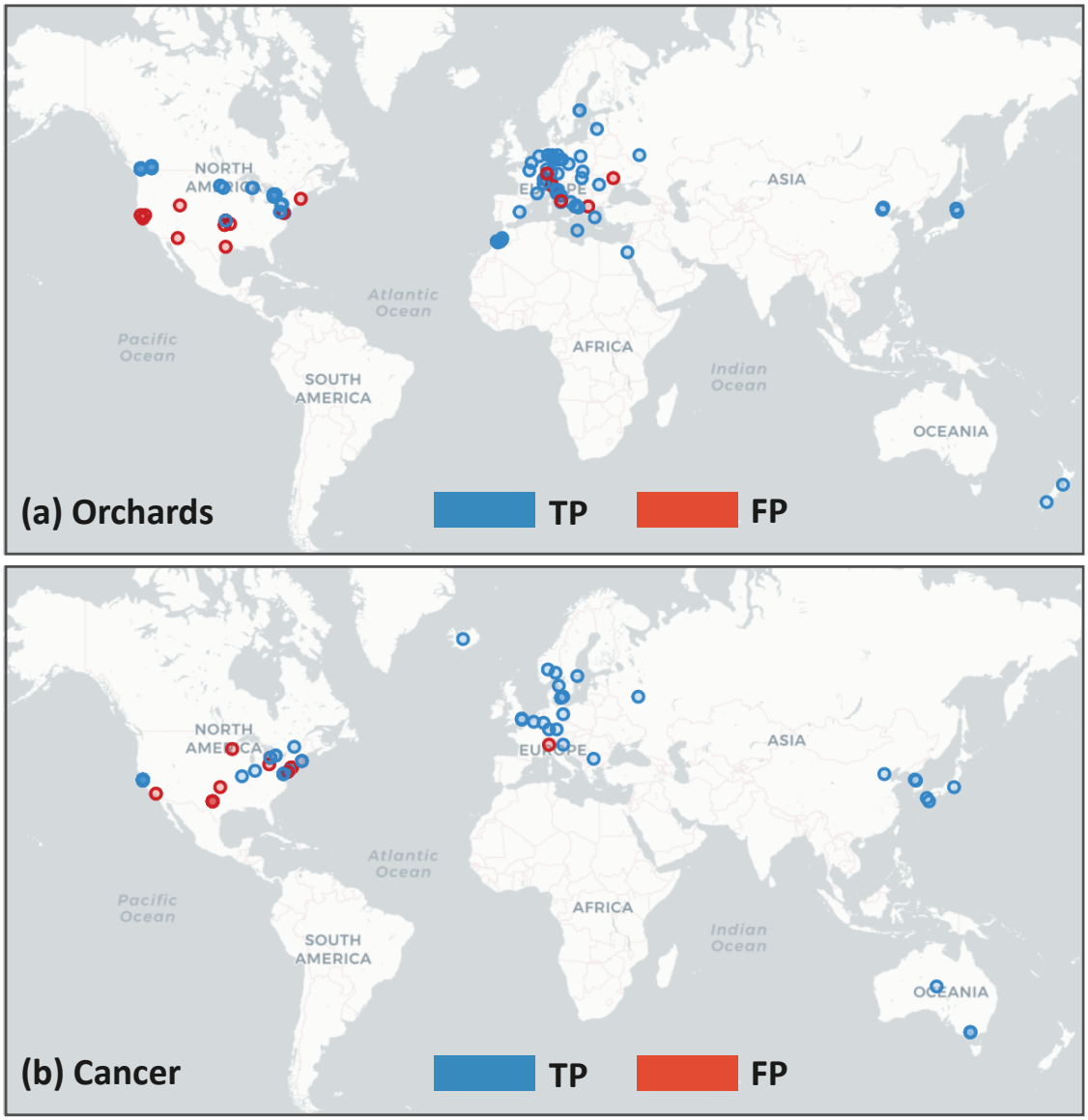


Figure 4.8: Maps of geocode results (true positives (TP) and false positives (FP)) for both test corpora. Figure from Acheson and Purves (submitted).

Summary

In this case study, we built a fully automatic text-to-space pipeline to extract relevant locations from scientific articles and obtained results of a high enough quality to be useful for a meta-analysis or to geographically search or filter articles ($F1$ of 0.81 for the Orchards corpus and 0.75 for the Cancer corpus). By using both an ecological and a biomedical corpus, we showed that it was possible to perform this task with limited domain-customization within the genre of scientific articles.

Our error analysis showed that errors in the final output were due to a multitude of causes, and hence pipeline improvements will likely result from improving various individual pipeline components. Writing a scientific paper is time-consuming and expensive and every scientific work could see its value increased via full-text analysis, an increasingly applicable solution thanks to open access policies.

Tell me what you think and then tell me what the really smart person in the room who disagrees with you thinks.

— Aaron Sorkin

5 Discussion

Contents

5.1 Revisiting the research gaps

We now revisit the research gaps we identified at the end of our overview of relevant literature (Section 2.6), identifying how each was addressed and where further progress could be made.

Spatial properties of global gazetteer resources

In Paper I (Acheson et al., 2017a), we conducted an in-depth quantitative analysis of two global gazetteers: GeoNames and TGN. We went beyond previous work, which had focused on global coverage and populated places in GeoNames (Graham and Sabbata, 2015), or looked at data quality issues in particular regions and countries in GeoNames (Ahlers, 2013), by considering a wider range of features types (all features, populated places, streams, mountains, hills) and doing so on the global scale. We developed a methodology to study data quality in global gazetteers, given the lack of

a high quality authoritative resource to use as ground truth, by comparing coverage and balance in two global gazetteers in a multi-scale, multi-method analysis.

Our results showed that the country unit is an especially important driver of coverage in these resources, with sharp changes in coverage observed over national borders, and a small set of countries in TGN with higher coverage than the rest, but overall low coverage in TGN, which agrees with previous findings (e.g. Ahlers, 2013). Furthermore, populated places generally had more consistent coverage across resources, while coverage of natural features was generally sparser and more idiosyncratic, a pattern of relative neglect of natural features consistent with previous work on natural feature representation quality (Mooney et al., 2010) and quantity (Bégin et al., 2013) in OSM. The immediate implication for text-to-space pipelines is that the coverage of any placename resource being considered in a task should be examined and evaluated for the region of interest and for the feature types of interest, where possible, and not assumed to be fit-for-purpose. Authoritative resources with defined data quality standards should generally be favored over resources which have amalgamated multiple other resources and hence have varying data quality over space.

Further work flowing from our analysis would include finding ways to remedy coverage and balance issues (for example, adjusting balance by subsetting data according to the coarsest granularity found across the desired region of interest), and finding ways to 'fill the gaps' in coverage (for example, by carefully merging complementary resources or by integrating crowdsourced content for low coverage areas). As for further analyzing coverage itself, it would be interesting to focus on the dynamic aspect of these resources by conducting a temporal analysis to study changes in coverage over time, including overnight changes from large data uploads.

Gazetteer matching methodology

In Paper II (Acheson et al., 2019), we developed a detailed gazetteer matching methodology, which we presented alongside a comprehensive review of existing work and a comparison of published methods and results. In addition, we publicly released

our code and annotated data in order for future work to build on our solutions. Our work dealt with an understudied subset of records, natural features, as opposed to the more widely studied populated places and POIs (Zheng et al., 2010; Martins, 2011; Dalvi et al., 2014; McKenzie et al., 2014). We showed how feature type information could be integrated into the matching process in a machine learning context using one-hot encoding when dealing with two different feature type hierarchies, rather than manually aligning type hierarchies (Hastings, 2008; Morana et al., 2014). Indeed, cross-gazetteer matching, as opposed to deduplication, likely involves dealing with multiple feature type hierarchies (Janowicz and Keßler, 2008), structural heterogeneity like differing schemas (Elmagarmid et al., 2007), and with gazetteers with different spatial properties, such as coverage and balance (Acheson et al., 2017a). Finally, we offered insight into the trade-offs between building a rule-based or machine-learning-based matching solution in an applied context, and pointed out important pitfalls to avoid when building a realistic machine learning pipeline.

Further work on gazetteer matching could apply our machine learning pipeline with random forests to a larger dataset, closer in size to the large POI datasets used in some previous work (Zheng et al., 2010; Dalvi et al., 2014). Any work on a novel dataset or region would however require new annotated data for training and evaluation. One interesting case study would be to match individual records in TGN to records in GeoNames over a multi-country region. This would both provide additional methodological challenges associated with the coverage patterns we documented in Paper I (Acheson et al., 2017a), and the results could provide a more detailed understanding of these coverage patterns based on the aligned records. For instance, it could be that TGN contains only a subset of 'important' records which are also in GeoNames, or TGN could provide, despite sparser coverage overall, a large subset of complementary records. Our gazetteer coverage analysis, by looking at record counts only, cannot make any statements about individual records and their alignment across the resources.

Natural feature types

In Paper I (Acheson et al., 2017a), we focused on data subsets of particular natural feature types (streams, mountains, and hills) in our analysis of coverage patterns in global gazetteers. We found that coverage of these natural features was sparser and more idiosyncratic than populated places in both GeoNames and TGN. In Paper II (Acheson et al., 2019), we again focused on natural feature types, this time in a gazetteer matching task, finding that matching performance on some types (such as streams, represented as points) was far lower than for other types (such as peaks), and that integrating feature types into the matching task (for example via one-hot encoding) was crucial, with all of our highest-performing models doing so. It would be interesting to consider more complex geometries than points for this matching task, since lines and polygons are available in swissNAMES3D for a large portion of records, including streams.

In Paper III (Wartmann et al., 2018), we built a text-to-space pipeline for hiking blogs, where many toponyms appearing in the texts were names of natural features. We benefited from the use of high quality placename resources (including the authoritative gazetteer for Switzerland, swissNAMES3D) to ground the manually identified toponyms. Future work could automatically identify placenames from these (or similar) German semi-formal narrative texts to see how well existing tools (in particular NER tools) perform on this particular subset of placenames. Work on similar texts (narrative texts in German with fine-granularity natural features) which involved identifying toponyms has been performed (Derungs and Purves, 2013), but performance on toponym recognition specifically is unclear. Future work on a corpus in a different region of the world, where high quality placename resources are not available, could test the importance of gazetteer quality on task performance. For a multi-country region, integrated or smoothed resources could be created and compared to unmodified gazetteers.

Textual data sources

We used non-traditional, understudied textual data sources in both of our case studies. In Paper III (Wartmann et al., 2018), we processed hiking blogs which are challenging because of the types of toponyms they contain (fine-grained entities like mountain huts and natural features like lakes and valleys) and the semi-formal language (where some toponyms are creatively spelled and vernacular names may be used). We obtained task-appropriate results using a combination of manual toponym recognition and automatic toponym grounding and filtering. Working on German-language texts also means dealing with potentially lower performing NER tools, which often first optimize their models for English, and dealing with particularities like de-casing toponyms (such as a genitive 's' in 'des Oeschinensees' or 'des Schwarzhorns'). Indeed, these two examples of language-related challenges could be the subject of further work, where automatic toponym recognition (placename identification) is required because of, for example, a larger corpus, and where explicit toponym disambiguation is required because of, for example, a lower overall toponym count per text.

In Paper IV (Acheson and Purves, submitted), we tackled a textual data source which is becoming an increasingly common object of study, but only rarely in the context of text-to-space pipelines: scientific articles. Our two corpora of scientific articles, for which we built a fully automatic text-to-space pipeline, presented many interesting challenges, including the need to deal with very imperfect NER output, and the need to heavily filter even correctly identified locations, due to the presence of many irrelevant, non-study site locations in these long texts. Furthermore, particularly in the Orchards corpus, many locations were compositional descriptions like '30km from Florence', which present their own set of recognition and grounding challenges. Though we made no attempt at extracting or interpreting these compositional descriptions, we did include cardinal directions and some spatial prepositions in our location 'chunks' in order to keep many words describing the same location together as one unit. In further work, our code could be adapted to recognize these types of expressions and these could be fed to a system similar

to the one described in van Erp et al. (2015). Unfortunately, current geocoding tools typically do not handle these expressions, despite long-standing calls to do so (Leidner and Lieberman, 2011), though some are currently working on this particular problem (Al-Olimat et al., 2019). Since not much would be immediately gained from parsing these types of expressions from our texts, we instead also provide the full sentence whenever a relevant location is identified.

Future work on scientific articles (or another dataset where only a subset of locations should be retained) could explore the use of sentence classification, including recently developed deep learning approaches (Kim, 2014), to identify a set of relevant locations by classifying each sentence in the article as either describing study/sample sites or not. Indeed, sentence classification has previously been used to classify sentences in scientific articles as 'environmental' (likely to contain study site information) or 'experimental' (likely to contain provenance information for chemicals) (Tamames and de Lorenzo, 2010). Our location extraction approaches could then be applied to only those sentences likely to contain study/sample site descriptions.

Geographical representation

In Paper III (Wartmann et al., 2018), our footprint generation process included point (placename) filtering using the centroid and standard deviation of all the candidate points, and our final footprints were in the form of both bounding boxes and convex hulls, created from the final filtered set of points. In an extension of the footprint generation process (Acheson et al., 2017c), we clustered candidate points (placenames) using DBSCAN, which in effect removes outliers and could substitute the filtering approach. Very few works have used clustering in the context of text-to-space pipelines, with notable exceptions including the use of DBSCAN for placename disambiguation in the context of reconstructing hiking trajectories (Moncla et al., 2014b), and a recent paper using clustering to disambiguate fine-grained places in the context of more urban place descriptions (Chen et al., 2019). As with our hiking blogs, these two case studies using clustering dealt with texts

featuring a suitably large number of textual mentions to fine-grained places in relatively close geographical proximity.

In Paper IV (Acheson and Purves, submitted), we also filtered candidate locations, but did so using rules prior to the geocoding step, and thus without making use of any geographical information such as their distribution in space. We output simple point geometries for all extracted locations, which is arguably appropriate for the global scale of our two article corpora, and consistent with previous works which use points to represent study sites (Wallis et al., 2011; Martin et al., 2012; Karl et al., 2013; Kmoch et al., 2018) or patient/virus locations (Weissenbacher et al., 2015; Tahsin et al., 2016). However, this is also clearly a limitation, and several authors acknowledge that other geometries would sometimes be better suited, such as polygons that capture the extent of a large study site (Wallis et al., 2011; Shapiro and Báldi, 2012; Karl et al., 2013; Karl, 2018). A range of options (a single point, a set of points, a bounding box, and a detailed polygon) are presented in Margulies et al. (2016), and further representational options could include circles (point-radius method) (Wieczorek et al., 2004) and probability density surfaces (Guo et al., 2008), both used to represent the location of natural history specimens. However, in an automatic process which uses a geocoding service, one is in practice limited to the geometries returned by the service, which are typically only points, though in our case a subset of results also contain a bounding box. Bounding boxes or circular regions could also be used to represent areal information efficiently, which could potentially be extracted from a subset of articles (particularly in the ecology domain).

Hill (2006) argues that one should capture and model uncertainty when representing locations in order to avoid erroneous conclusions due to 'false precision': "The level of precision that is 'right' is judged not only by the level of confidence in the data but also by the detail needed for a particular purpose and for foreseen future uses." Hill's recommendations include documenting known uncertainties about footprints in gazetteers, storing multiple footprints per gazetteer entry, and making representational choices for footprints based not only on current uses, but

also anticipated uses. In this spirit, our outputs include locations represented in both textual and explicitly spatial form, such that spatial representations can be re-generated and refined as needed using the location strings.

Real-world motivated case studies

Our two case studies were motivated by real-world needs. In Paper III (Wartmann et al., 2018), the requirements for our text-to-space pipeline stemmed from an immediate downstream task - to spatially query geotagged photos (points) within our generated study site footprints (polygons) - which itself was part of a broader study aiming to collect and compare landscape descriptions from the public for future use in landscape policy. In Paper IV (Acheson and Purves, submitted), we automated the study site location or sample location extraction process which two collaborators working on separate meta-analyses were carrying out manually: one by looking through the article contents for the relevant information, and the other by using the first author location as a proxy. The text-to-space pipeline we built is tailored to scientific articles, but since we limited and isolated our domain-specific code customizations, it should be applicable to new corpora with low additional effort. In both of our case studies, the ultimate purpose of the text-to-space pipeline helped guide the many processing decisions that needed to be made during development and helped determine what acceptable performance meant.

In Paper II (Acheson et al., 2019), we built a real-world driven machine learning pipeline for gazetteer matching. Indeed, we took great care to ensure the processing decisions applied to our annotated data would reflect, and be applicable to, a real-world situation, where a pipeline needs to be deployed to new, unannotated data. This meant, for example, that candidate selection for our test data had to be 'naive' about what the true positive pairs were, in order to get a more realistic full pipeline performance estimate. In Paper I (Acheson et al., 2017a), we focused primarily on the GeoNames gazetteer due to it being by far the most commonly used gazetteer in the broad literature read in the context of this thesis.

This means our gazetteer comparison and analysis should be valuable to more people, within and outwith academia.

Reproducibility

To maximize the reproducibility of the work done in the context of this thesis, we published our code alongside releasable data for Paper II (Acheson et al., 2019) and Paper IV (Acheson and Purves, submitted). Anyone wishing to use or extend our work can thus freely access our code alongside high-level instructions and more detailed documentation *in situ* in the code. Earlier efforts to further reproducible research in GIScience and Geography resulted in the co-organization of a workshop of which the proceedings are hosted here: `http://www.geo.uzh.ch/microsite/reproducible_research/`.

5.2 Customizing a text-to-space pipeline

Throughout this thesis, various claims are made that certain resources, tools, and strategies are more or less suitable for a particular use case, depending on the task requirements or the properties of the text corpora. Though it is not an aim of this thesis to produce an 'instruction manual' on how to build a text-to-space pipeline for a new use case, several insights are worth considering for future pipeline development:

- **Textual corpus properties:** The language of the texts and an estimate of how formal the texts are will help decide on an initial placename identification (toponym recognition) strategy, where an out-of-the-box NER tool like Stanford NER can be expected to perform well on formal text in a supported language (for Stanford NER: English, German, Spanish, and Chinese[1]). As the texts deviate in form and content from the type of texts used to train out-of-the-box NER tools, lower performance should be expected, and NER tools could be retrained on an annotated portion of the particular dataset to be processed.

[1] As documented here: https://nlp.stanford.edu/software/CRF-NER.html (accessed in 07.2019)

- **Spatial corpus properties:** Some knowledge about the geographical entities that are mentioned in text should be acquired at the start of a project. Questions include: What is the general region of the entities (global or within a particular country)? Do we expect locations within a text to be clustered in space? Are the entities fine-grained (urban POIs, mountain huts, etc.) or rather coarse-grained (cities, national parks, etc.)? What specific feature types predominate (for example, natural or human-made entities)? Answers to these questions should help select adequate placenames resources (ones which cover the correct region and types at an appropriate level of detail) and disambiguation strategies (with distance-based strategies, like filtering and clustering, suitable when there is a meaningful core region to be identified).
- **Spatio-textual corpus properties:** The quantity, density, and location of placenames in text also play a role in the success of placename identification and grounding strategies, though we have not explored this particular issue much in this thesis. A text where a high proportion of the total word count consist of placenames may benefit from re-trained NER tools that can consider this density, or potentially a simpler location identification strategy such as gazetteer lookup; conversely, a text with a low proportion of placenames may require careful examination of the NER output to filter false positives. Some texts, like news articles, may feature important, overarching locations in the title or in the opening sentence(s) (Alex and Grover, 2010). Our Orchards scientific articles corpora had key locations mentioned in the title in more than a third of the articles, but our Cancer corpus had virtually none.
- **Representational task requirements:** For some tasks, points may be an appropriate way to represent all entities, for example if dealing with cities and finer-grained entities on a global scale. In other cases, areas are necessary, for example to perform point-in-polygon spatial queries. For these, polygonal representations for individual entities could be obtained (for example, OSM Nominatim can return geojson polygons for many entities), or areas can be created from a set of points (Galton and Duckham, 2006).

5.3 Limitations and perspectives

Though we touched upon limitations and future directions related to specific research gaps in the previous section, several broader limitations and corresponding future directions are worth adding:

- In our case studies, we did not systematically compare a pre-defined set of processing options (such as text pre-processing, NER tools, and so on) for the various steps involved. Though in the previous section we suggest ways of making intelligent guesses to customize a text-to-space pipeline, a more systematic way to optimize a pipeline would be desirable and could be the subject of future work. However, the dynamic landscape of tools and resources makes this particularly challenging. In related work, some encouraging progress is being made towards comparing the performance of various toponym identification and resolution systems on various open datasets (Gritta et al., 2017; Hu, 2018).
- Continuing with the theme of systematic evaluation, we also did not quantitatively measure the influence of resources on task performance. Future work could look at how sensitive results are, for a specified task, to changes in the resources used (particularly gazetteers, but also NER tools or other placename identification and disambiguation tools). Our gazetteer coverage analysis and case studies suggest, but don't quantitatively measure, this influence. Though we tested and compared NER tools and geocoders, for example on our scientific articles datasets, we did so privately in an ad hoc way on data subsets, but again here more systematic benchmarking and evaluation would be desirable, particularly ones involving multiple independent research groups under time constraints.
- In general, our text-to-space pipelines and our gazetteer matching code could be more efficient. As we worked on relatively small datasets, it was not a requirement to process large datasets in a small amount of time and hence efficiency often took a backseat to task-performance metrics like precision and

recall. In gazetteer matching, some candidate selection strategies in particular could be both more efficient and more flexible, and in our scientific articles processing, the file conversion and NER steps are particularly slow, though each article gets processed independently of the others, hence articles could be processed in parallel.

- We limited ourselves to point geometries to represent individual places. Though many gazetteers and geocoders offer point-based geometries to represent all or most named places, polygons for countries are freely available from various sources. It is also possible to obtain polygons for many entities from a geocoder like OSM Nominatim, which can return polygons when this option is specified. In our case studies, using points to represent individual entities satisfied our requirements. However, it would be interesting to work on a spatially-rich, mixed-granularity corpus where spatial relationships between entities, like figure-ground or container-contained, could be detected and depicted. Regional and international travel blogs would be an interesting data source for this type of work, where large and small entities would likely mix, including countries, cities, POIs, and natural features, and additionally itineraries could be depicted (essentially temporal relationships). Some interesting work which considered itineraries has been done on fine-granularity hiking blogs (Moncla et al., 2014a) and travel blogs have been used as a data source in, for example, the geographically-aware exploratory search system Frankenplace (Adams et al., 2015).
- Continuing with the theme of spatial relationships between entities, we did not explicitly take into consideration spatial prepositions and the surrounding textual context of placenames (such as in the case of compositional descriptions) for place or document modeling. Indeed, another way to detect and model spatial relationships between entities would be to start from the text itself, rather than from external information about the entities. Studying spatial relationships could focus on the use of spatial prepositions to encode spatial relationships in language, in a geographical context (as opposed to,

for example, in the context of 3D manipulable object space). This should also involve studying the nature of the entities themselves, including their feature type and size (Hall et al., 2015; Stock and Yousaf, 2018).

Life is too short to occupy oneself with the slaying of the slain more than once.

— Thomas Henry Huxley

6 Conclusions

Contents

6.1 Summary and contributions

This thesis presented work relating to text-to-space pipelines, focusing both on placename *resources* (gazetteers) and on specific *applications* (case studies). A concise overview of relevant literature was presented in Chapter 2, building up to a set of research gaps that we sought to address in our papers. Chapter 3 described work on placename resources, summarizing methods and results from Paper I (Acheson et al., 2017a) on the spatial coverage of global gazetteers and Paper II (Acheson et al., 2019) on cross-gazetteer matching. Chapter 4 described the two case studies where text-to-space pipelines were implemented for particular applications, summarizing methods and results from Paper III (Wartmann et al., 2018) on generating footprints from hiking blogs and Paper IV (submitted) on extracting and representing geographical information from scientific articles. In Chapter 5, we returned to the research gaps identified early in this thesis, describing

how each was filled and where additional efforts could be focused, and we also offered more general insights into building text-to-space pipelines based on our experience.

Our main contributions can be summarized as follows:

- We quantitatively analyzed and compared the spatial coverage of two global gazetteers, GeoNames and TGN, looking at important subsets of records including common natural feature types. Our analysis showed large differences in the amount and type of records contained in these resources depending on the region, and identified several patterns of coverage, including the country unit as an important driver of coverage. We discussed how the spatial coverage of gazetteers has implications for the performance of text-to-space tasks which make use of gazetteers, including identifying toponyms in text and linking toponyms to particular gazetteer records, since only records present in a gazetteer can be linked to, and regions with high spatial coverage may get overrepresented in task results (and vice-versa).
- We proposed, implemented, and conducted a detailed evaluation of a machine learning pipeline to match natural feature records across two gazetteers. We compared the performance of machine-learning based matching using random forests to rule-based matching, showing that random forests performed better than our best rules, offered potentially further increases in performance given more training data, and simplified the handling of feature types from different feature type hierarchies.
- We implemented a text-to-space pipeline to generate spatial footprints from Swiss hiking blogs which featured many fine-granularity natural feature toponyms. Through our application to the analysis of landscape descriptions from multiple data sources, we showed how such a text-to-space pipeline could be used to spatially query other georeferenced data sources (such as social media). We showed how outliers could be removed to obtain focused footprints without explicitly disambiguating each placename, but instead by using centroid-based filtering and DBSCAN clustering of candidate placenames/points.

- We implemented a fully-automatic text-to-space pipeline to extract study sites and patient treatment locations from scientific articles in two different corpora. We showed that good performance could be obtained to extract and spatially represent relevant locations from these texts by combining existing NER and geocoding tools with rules developed iteratively on a set of training articles. We also showed that minimal domain-customization was needed across our two domains (biomedical and ecological), within the genre of scientific articles. Our geocoding results suggested that the black-box geocoding API we used tended to default to results in North America, based on the geographical distribution of errors.

6.2 Future directions

The work in this thesis could be extended in many directions within the field of GIScience, including by:

- Parsing, grounding, and representing not just toponyms but a wider variety of spatial language and context including the prepositions used in conjunction with toponyms, and relationships between places mentioned in a discourse.
- Using richer geographic models for individual locations and testing ways to usefully combine many geometries to represent a document as a whole, or a set of related documents.
- Modeling locations to a granularity, crispness, precision, or certainty justified by the data, that is, with effort put into matching the likely cognitive models of the author.
- Doing user testing on the visualization of, and interaction with, different representations of individual geographic entities of various types, individual text documents from various genres and domains, and corpus-level representations, ideally all in the context of specific tasks.

Bibliography

Acheson, E., De Sabbata, S., and Purves, R. S. (2017a). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320.

Acheson, E., Villette, J., Volpi, M., and Purves, R. S. (2017b). Gazetteer Matching for Natural Features in Switzerland. In *Proceedings of the 11th Workshop on Geographic Information Retrieval*, GIR'17, pages 11:1–11:2, New York, NY, USA. ACM.

Acheson, E., Volpi, M., and Purves, R. S. (2019). Machine learning for cross-gazetteer matching of natural features. *International Journal of Geographical Information Science*, pages 1–27.

Acheson, E., Wartmann, F. M., and Purves, R. S. (2017c). Generating Spatial Footprints from Hiking Blogs. In *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*, Lecture Notes in Geoinformation and Cartography, pages 5–7. Springer, Cham.

Adams, B., McKenzie, G., and Gahegan, M. (2015). Frankenplace: Interactive Thematic Mapping for Ad Hoc Exploratory Search. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 12–22, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Ahlers, D. (2013). Assessment of the Accuracy of GeoNames Gazetteer Data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, pages 74–81, New York, NY, USA. ACM.

Al-Olimat, H. S., Shalin, V. L., Thirunarayan, K., and Sain, J. P. (2019). Towards Geocoding Spatial Expressions. *arXiv:1906.04960 [cs]*. arXiv: 1906.04960.

Alex, B. and Grover, C. (2010). Labelling and Spatio-temporal Grounding of News Events. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 27–28, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alex, B., Grover, C., Oberlander, J., Thomson, T., Anderson, M., Loxley, J., Hinrichs, U., and Zhou, K. (2017). Palimpsest: Improving assisted curation of loco-specific literature. *Digital Scholarship in the Humanities*, 32(suppl_1):i4–i16.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: Geotagging Web Content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 273–280, New York, NY, USA. ACM.

Anastácio, I., Martins, B., and Calado, P. (2009). A Comparison of Different Approaches for Assigning Geographic Scopes to Documents. In *In Proceedings of the 1st INForum - Simpósio de Informática*.

ARE (2011). Landschaftstypologie Schweiz Teil 1, Ziele, Methode und Anwendung. Technical report, ARE, Berne, Switzerland.

Augenstein, I., Derczynski, L., and Bontcheva, K. (2017). Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Axelrod, A. E. (2003). On Building a High Performance Gazetteer Database. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*, HLT-NAACL-GEOREF '03, pages 63–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC 2004*, pages 1313–1316, Lisbon, Portugal. ELRA.

Batista, D. S., Silva, M. J., Couto, F. M., and Behera, B. (2010). Geographic Signatures for Semantic Retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 19:1–19:8, New York, NY, USA. ACM.

Bennett, B. and Agarwal, P. (2007). Semantic categories underlying the meaning of 'place'. In *Spatial information theory*, pages 78–95. Springer.

Bégin, D., Devillers, R., and Roche, S. (2013). Assessing volunteered geographic information (VGI) quality based on contributors' mapping behaviours. In *Proceedings of the 8th international symposium on spatial data quality ISSDQ*, pages 149–154.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Bordogna, G., Ghisalberti, G., and Psaila, G. (2012). Geographic information retrieval: Modeling uncertainty of user's context. *Fuzzy Sets and Systems*, 196:105–124.

Brando, C., Dominguès, C., and Capeyron, M. (2016). Evaluation of NER systems for the recognition of place mentions in French thematic corpora. In *Proceedings of the 10th Workshop on Geographic Information Retrieval*, San Francisco.

Brauner, D. F., Casanova, M. A., and Milidiú, R. L. (2007). Towards Gazetteer Integration through an Instance-based Thesauri Mapping Approach. In Jr, C. A. D. and Monteiro, A. M. V., editors, *Advances in Geoinformatics*, pages 235–245. Springer Berlin Heidelberg.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brunner, T. J. and Purves, R. S. (2008). Spatial Autocorrelation and Toponym Ambiguity. In *Proceedings of the 5th Workshop on Geographic Information Retrieval*, GIR '08, pages 25–26, New York, NY, USA. ACM.

Budig, B. and van Dijk, T. C. (2017). Journeys of the Past: A Hidden Markov Approach to Georeferencing Historical Itineraries. In *Proceedings of the 11th Workshop on Geographic Information Retrieval*, GIR'17, pages 7:1–7:10, New York, NY, USA. ACM. event-place: Heidelberg, Germany.

Burenhult, N. and Levinson, S. C. (2008). Language and landscape: a cross-linguistic perspective. *Language Sciences*, 30(2–3):135–150.

Burrough, P. A. and Frank, A. (1996). *Geographic Objects with Indeterminate Boundaries*. CRC Press.

Buscaldi, D. (2011). Approaches to Disambiguating Toponyms. *SIGSPATIAL Special*, 3(2):16–19.

Buyukokkten, O., Cho, J., Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1999). Exploiting Geographical Location Information of Web Pages. Philadelphia, Pennsylvania.

Cai, H., Kumar, N., Ai, N., Gupta, S., Rath, P., and Baudis, M. (2014). Progenetix: 12 years of oncogenomic data curation. *Nucleic Acids Research*, 42(Database issue):D1055–1062.

Chen, H., Vasardani, M., and Winter, S. (2019). Clustering-based disambiguation of fine-grained place names from descriptions. *GeoInformatica*.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA. ACM.

Christen, P. (2012). A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1537–1555.

Clough, P. (2005). Extracting Metadata for Spatially-aware Information Retrieval on the Internet. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, GIR '05, pages 25–30, New York, NY, USA. ACM.

Coates, R. (2006). Properhood. *Language*, 82(2):356–382.

Costa, S., Ogilvie, D., Dalton, A., Westgate, K., Brage, S., and Panter, J. (2015). Quantifying the physical activity energy expenditure of commuters using a combination of global positioning system and combined heart rate and movement sensors. *Preventive Medicine*.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., Karagiorgou, S., Efentakis, A., and Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, 29(5):720–741.

Dalvi, N., Olteanu, M., Raghavan, M., and Bohannon, P. (2014). Deduplicating a Places Database. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 409–418, Seoul, Korea. ACM.

De Sabbata, S. and Acheson, E. (2016). Geographies of gazetteers in Great Britain. In *24th GIS Research UK (GISRUK 2016) conference*, Greenwich, UK.

DeLozier, G., Baldridge, J., and London, L. (2015). Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Derungs, C. (2014). *From Text to Landscape: Extraction of Landscape Concepts through the Resolution of Ambiguity and Vagueness present in Descriptions of Natural Landscapes.* PhD thesis.

Derungs, C. and Purves, R. S. (2013). From text to landscape: locating, identifying and

mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, 28(6):1272–1293.

Dias, D., Anastácio, I., and Martins, B. (2012). A Language Modeling Approach for Georeferencing Textual Documents. *Proceedings of the 2nd Spanish Conference in Information Retrieval*.

Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing Geographical Scopes of Web Resources. In *Proceedings of the 26th VLDB Conference*, Cairo, Egypt.

Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, pages 20–24. Citeseer.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Trans. on Knowl. and Data Eng.*, 19(1):1–16.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., and others (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics. event-place: Ann Arbor, Michigan.

Fisher, R., Radford, B. T., Knowlton, N., Brainard, R. E., Michaelis, F. B., and Caley, M. J. (2011). Global mismatch between research effort and conservation needs of tropical coral reefs. *Conservation Letters*, 4(1):64–72.

Fu, G., Jones, C. B., and Abdelmoty, A. I. (2005). Building a Geographical Ontology for Intelligent Spatial Search on the Web. In *Proceedings of IASTED International Conference on Databases and Applications (DBA-2005)*, pages 167–172, Innsbruck, Austria. ACTA Press.

Galton, A. and Duckham, M. (2006). What is the region occupied by a set of points? In *Geographic Information Science*, pages 81–98. Springer.

Gan, Q., Attenberg, J., Markowetz, A., and Suel, T. (2008). Analysis of Geographic Queries in a Search Engine Log. In *Proceedings of the First International Workshop on Location and the Web*, LOCWEB '08, pages 49–56, New York, NY, USA. ACM.

Gao, S., Li, L., Li, W., Janowicz, K., and Zhang, Y. (2017). Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*, 61, Part B:172–186.

Gelernter, J., Ganesh, G., Krishnakumar, H., and Zhang, W. (2013). Automatic Gazetteer Enrichment with User-geocoded Data. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GEOCROWD '13, pages 87–94, New York, NY, USA. ACM.

Gelernter, J. and Zhang, W. (2013). Cross-lingual Geo-parsing for Non-structured Data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, pages 64–71, New York, NY, USA. ACM. event-place: Orlando, Florida.

Gonçalves, N. F. A. (2012). Gazetteer Record Linkage. Master's thesis, Instituto Superior Técnico, Lisbon.

Graham, M. and Sabbata, S. D. (2015). Mapping information wealth and poverty: the geography of gazetteers. *Environment and Planning A*, 47(6):1254–1264.

Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2017). What's missing in geographical parsing? *Language Resources and Evaluation*, pages 1–21.

Guo, Q., Liu, Y., and Wieczorek, J. (2008). Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10):1067–1090.

Guptill, S. C. and Morrison, J. L. (1995). *Elements of Spatial Data Quality*. Elsevier Science Limited.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4):682–703.

Hall, M. M., Jones, C. B., and Smart, P. (2015). Spatial Natural Language Generation for Location Description in Photo Captions. In Fabrikant, S. I., Raubal, M., Bertolotto, M., Davies, C., Freundschuh, S., and Bell, S., editors, *Spatial Information Theory*, volume 9368, pages 196–223, Cham. Springer International Publishing.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pages 451–500.

Hastings, J. T. (2008). Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, 22(10):1109–1127.

Herskovits, A. (1985). Semantics and Pragmatics of Locative Expressions*. *Cognitive Science*, 9(3):341–378.

Hess, B., Magagna, F., and Sutanto, J. (2014). Toward location-aware Web: extraction method, applications and evaluation. *Personal and Ubiquitous Computing*, 18(5):1047–1060.

Hill, L. L. (2000). Core elements of digital gazetteers: placenames, categories, and footprints. In *Research and advanced technology for digital libraries*, pages 280–290. Springer.

Hill, L. L. (2006). *Georeferencing: The Geographic Associations of Information*. The MIT Press.

Hollenstein, L. and Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 0(1):21–48.

Hu, Y. (2018). EUPEG: Towards an Extensible and Unified Platform for Evaluating Geoparsers. In *Proceedings of the 12th Workshop on Geographic Information Retrieval - GIR'18*, pages 1–2, Seattle, WA, USA. ACM Press.

Janowicz, K. and Keßler, C. (2008). The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 22(10):1129–1157.

Jiang, R., Banchs, R. E., and Li, H. (2016). Evaluating and Combining Named Entity Recognition Systems. *ACL 2016*, page 21.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., and Wallgrün, J. O. (2018). GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 0(0).

Karl, J. W. (2018). Mining location information from life- and earth-sciences studies to facilitate knowledge discovery. *Journal of Librarianship and Information Science*, page 0961000618759413.

Karl, J. W., Herrick, J. E., Unnasch, R. S., Gillan, J. K., Ellis, E. C., Lutters, W. G., and Martin, L. J. (2013). Discovering Ecologically Relevant Knowledge from Published Studies through Geosemantic Searching. *BioScience*, 63(8):674–682.

Kelleher, J. D. and Costello, F. J. (2008). Applying Computational Models of Spatial Prepositions to Visually Situated Dialog. *Computational Linguistics*, 35(2):271–306.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kmoch, A., Uuemaa, E., Klug, H., and Cameron, S. G. (2018). Enhancing Location-Related Hydrogeological Knowledge. *ISPRS International Journal of Geo-Information*, 7(4):132.

Larson, R. R. (1996). *Geographic information retrieval and spatial browsing*. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Leidner, J. L. (2004a). Toponym resolution in text:"Which Sheffield is it?". In *Proceedings of the the 27th annual international ACM SIGIR conference (SIGIR 2004)*, page 602. Citeseer.

Leidner, J. L. (2004b). Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on Geographic Information Retrieval, Sheffield, UK*.

Leidner, J. L. (2007). *Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names*. PhD thesis, Edinburgh University, Edinburgh, Scotlan.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11.

Leveling, J. (2015). Tagging of Temporal Expressions and Geological Features in Scientific Articles. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, GIR '15, pages 6:1–6:10, New York, NY, USA. ACM.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Lieberman, M. D., Samet, H., Sankaranarayanan, J., and Sperling, J. (2007). STEWARD: Architecture of a Spatio-textual Search Engine. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*, GIS '07, pages 25:1–25:8, New York, NY, USA. ACM.

Liu, F., Vasardani, M., and Baldwin, T. (2014). Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis. In *Proceedings of the 4th International Workshop on Location and the Web*, LocWeb '14, pages 9–16, New York, NY, USA. ACM.

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (2005). *Geographic Information Systems and Science*. John Wiley & Sons. Google-Books-ID: toobg6OwFPEC.

Maaß, W., Baus, J., and Paul, J. (1995). *Visual grounding of route descriptions in dynamic environments*. Univ. des Saarlandes, SFB 314.

Magge, A., Weissenbacher, D., Sarker, A., Scotch, M., and Gonzalez-Hernandez, G. (2018). Bi-directional Recurrent Neural Network Models for Geographic Location Extraction in Biomedical Literature. In *Biocomputing 2019*, pages 100–111. WORLD SCIENTIFIC.

Mahmud, J., Nichols, J., and Drews, C. (2012). Where Is This Tweet From? Inferring Home Locations of Twitter Users. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Margulies, J. D., Magliocca, N. R., Schmill, M. D., and Ellis, E. C. (2016). Ambiguous Geographies: Connecting Case Study Knowledge with Global Change Science. *Annals of the American Association of Geographers*, 106(3):572–596.

Markowetz, A., Chen, Y.-Y., Suel, T., Long, X., and Seeger, B. (2005). Design and Implementation of a Geographic Search Engine. In *WebDB*, volume 2005, pages 19–24.

Martin, L. J., Blossey, B., and Ellis, E. (2012). Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment*, 10(4):195–201.

Martins, B. (2011). A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records. In *GeoSpatial Semantics*, Lecture Notes in Computer Science, pages 34–51. Springer, Berlin, Heidelberg.

Martins, B. and Silva, M. J. (2005). A graph-ranking algorithm for geo-referencing documents. In *2013 IEEE 13th International Conference on Data Mining*, pages 741–744. IEEE Computer Society.

McKenzie, G., Janowicz, K., and Adams, B. (2014). A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science*, 41(2):125–137.

Middleton, S. E., Middleton, L., and Modafferi, S. (2014). Real-Time Crisis Mapping of Natural Disasters Using Social Media. *IEEE Intelligent Systems*, 29(2):9–17.

Moncla, L., Gaio, M., and Mustière, S. (2014a). Automatic Itinerary Reconstruction

from Texts. In *Automatic Itinerary Reconstruction from Texts*, volume 8728, pages pp.253–267, Vienna, Austria. Springer International Publishing.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., and Gaio, M. (2014b). Geocoding for Texts with Fine-grain Toponyms: An Experiment on a Geoparsed Hiking Descriptions Corpus. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 183–192, Dallas/Fort Worth, TX, USA. ACM.

Montello, D. R. (1993). Scale and multiple psychologies of space. In Frank, A. U. and Campari, I., editors, *Spatial Information Theory A Theoretical Basis for GIS*, number 716 in Lecture Notes in Computer Science, pages 312–321. Springer Berlin Heidelberg.

Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P. (2003). Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition & Computation*, 3(2-3):185–204.

Mooney, P., Corcoran, P., and Winstanley, A. C. (2010). A study of data representation of natural features in OpenStreetMap. In *Proceedings of GIScience*, volume 150.

Morana, A., Morel, T., Berjawi, B., and Duchateau, F. (2014). GeoBench: A Geospatial Integration Tool for Building a Spatial Entity Matching Benchmark. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, pages 533–536, New York, NY, USA. ACM.

Murrieta-Flores, P., Baron, A., Gregory, I., Hardie, A., and Rayson, P. (2015). Automatically Analyzing Large Texts in a GIS Environment: The Registrar General's Reports and Cholera in the 19th Century. *Transactions in GIS*, 19(2):296–320.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Overell, S. and Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22(3):265–287.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, d. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227.

Pinker, S. (2008). *The Stuff of Thought: Language as a Window into Human Nature*. Penguin UK. Google-Books-ID: 3DocCGB0cRkC.

Popescu, A., Grefenstette, G., and Moëllic, P. A. (2008). Gazetiki: Automatic Creation of a Geographical Gazetteer. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, pages 85–93, New York, NY, USA. ACM.

Purves, R. and Jones, C. (2011). Geographic Information Retrieval. *SIGSPATIAL Special*, 3(2):2–4.

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A. K., Vaid, S., and Yang, B. (2007). The design and implementation

of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7):717–745.

Rahimi, A., Cohn, T., and Baldwin, T. (2016). pigeo: A python geotagging tool. *Proceedings of ACL-2016 System Demonstrations*, pages 127–132.

Rauch, E., Bukatin, M., and Baker, K. (2003). A Confidence-based Framework for Disambiguating Geographic Terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*, HLT-NAACL-GEOREF '03, pages 50–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reuschel, A.-K. and Hurni, L. (2011). Mapping Literature: Visualisation of Spatial Uncertainty in Fiction. *The Cartographic Journal*, 48(4):293–308.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. (2012). Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510.

Sehgal, V., Getoor, L., and Viechnicki, P. D. (2006). Entity Resolution in Geospatial Data Integration. In *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems*, GIS '06, pages 83–90, New York, NY, USA. ACM.

Shapiro, J. T. and Báldi, A. (2012). Lost locations and the (ir)repeatability of ecological studies. *Frontiers in Ecology and the Environment*, 10(5):235–236.

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., and Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378–399.

Smart, P. D., Jones, C. B., and Twaroch, F. A. (2010). Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service. In *Geographic Information Science*, Lecture Notes in Computer Science, pages 234–248. Springer, Berlin, Heidelberg.

Smith, D. A. and Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library. In Constantopoulos, P. and Sølvberg, I. T., editors, *Research and Advanced Technology for Digital Libraries*, number 2163 in Lecture Notes in Computer Science, pages 127–136. Springer Berlin Heidelberg.

Smith, D. A. and Mann, G. S. (2003). Bootstrapping Toponym Classifiers. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*, HLT-NAACL-GEOREF '03, pages 45–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stock, K. and Yousaf, J. (2018). Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. *International Journal of Geographical Information Science*, 0(0):1–30.

Tahsin, T., Weissenbacher, D., Rivera, R., Beard, R., Firago, M., Wallstrom, G., Scotch, M., and Gonzalez, G. (2016). A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records. *Journal of the American Medical Informatics Association*, 23(5):934–941.

Talmy, L. (1983). How language structures space. In *Spatial Orientation: Theory, Research, and Application*, pages 225–282.

Tamames, J. and de Lorenzo, V. (2010). EnvMine: A text-mining system for the automatic extraction of contextual information. *BMC Bioinformatics*, 11:294.

Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. (2008). NewsStand: A New View on News. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, pages 18:1–18:10, New York, NY, USA. ACM.

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., and Bolikowski, L. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335.

van der Meer, M., Lüscher, G., Kay, S., and Jeanneret, P. (2017). What evidence exists on the impact of agricultural practices in fruit orchards on biodiversity indicator species groups? A systematic map protocol. *Environmental Evidence*, 6:14.

van Erp, M., Hensel, R., Ceolin, D., and Meij, M. v. d. (2015). Georeferencing Animal Specimen Datasets. *Transactions in GIS*, 19(4):563–581.

Van Laere, O., Schockaert, S., and Dhoedt, B. (2013). Georeferencing Flickr resources based on textual meta-data. *Information Sciences*, 238:52–74.

Van Laere, O., Schockaert, S., Tanasescu, V., Dhoedt, B., and Jones, C. B. (2014). Georeferencing Wikipedia Documents Using Data from Social Media Sources. *ACM Trans. Inf. Syst.*, 32(3):12:1–12:32.

Vasardani, M., Winter, S., and Richter, K.-F. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27(12):2509–2532.

Vincenty, T. (1975). Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. *Survey Review*, 23(176):88–93.

Wallis, P. J., Nally, R. M., and Langford, J. (2011). Mapping Local-Scale Ecological Research to Aid Management at Landscape Scales: Mapping Ecological Research at Landscape Scales. *Geographical Research*, 49(2):203–216.

Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W.-Y. (2005). Detecting Geographic Locations from Web Resources. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, GIR '05, pages 17–24, New York, NY, USA. ACM.

Wartmann, F. M., Acheson, E., and Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach. *International Journal of Geographical Information Science*, 32(8):1572–1592.

Wartmann, F. M. and Purves, R. S. (2018). Investigating sense of place as a cultural ecosystem service in different landscapes through the lens of language. *Landscape and Urban Planning*, 175:169–183.

Weissenbacher, D., Magge, A., O'Connor, K., Scotch, M., and Gonzalez, G. (2019). SemEval-2019 Task 12: Toponym Resolution in Scientific Papers. In *Proceedings of the*

13th International Workshop on Semantic Evaluation (SemEval-2019), pages 907–916, Minneapolis, Minnesota, USA.

Weissenbacher, D., Sarker, A., Tahsin, T., Scotch, M., and Gonzalez, G. (2017). Extracting geographic locations from the literature for virus phylogeography using supervised and distant supervision methods. *AMIA Summits on Translational Science Proceedings*, 2017:114–122.

Weissenbacher, D., Tahsin, T., Beard, R., Figaro, M., Rivera, R., Scotch, M., and Gonzalez, G. (2015). Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. *Bioinformatics*, 31(12):i348–i356.

Wieczorek, J., Guo, Q., and Hijmans, R. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18(8):745–767.

Wing, B. P. and Baldridge, J. (2011). Simple Supervised Document Geolocation with Geodesic Grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 955–964, Stroudsburg, PA, USA. Association for Computational Linguistics.

Won, M., Murrieta-Flores, P., and Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5.

Woodruff, A. G. and Plaunt, C. (1994). GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655.

Worboys, M. F. (2001). Nearness relations in environmental space. *International Journal of Geographical Information Science*, 15(7):633–651.

Yin, J., Karimi, S., and Lingad, J. (2014). Pinpointing Locational Focus in Microblogs. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, pages 66:66–66:72, New York, NY, USA. ACM.

Zhang, W. and Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, (9).

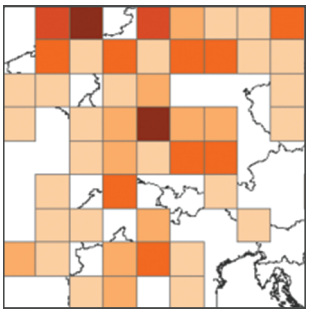
Zheng, Y., Fen, X., Xie, X., Peng, S., and Fu, J. (2010). Detecting Nearly Duplicated Records in Location Datasets. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 137–143, New York, NY, USA. ACM.

Zhu, R., Hu, Y., Janowicz, K., and McKenzie, G. (2016). Spatial Signatures for Geographic Feature Types: Examining Gazetteer Ontologies using Spatial Statistics. *Transactions in GIS*.

Zong, W., Wu, D., Sun, A., Lim, E.-P., and Goh, D. H.-L. (2005). On Assigning Place Names to Geography Related Web Pages. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '05, pages 354–362, New York, NY, USA. ACM.

Part II

Papers



Paper I

A quantitative analysis of global gazetteers: Patterns of coverage for common feature types

Summary

This paper analyzes and compares the spatial coverage of two global gazetteers, GeoNames and the Getty Thesaurus of Geographic Names (TGN), using point density maps, correlations, and linear regressions. The full datasets are analyzed, as well as feature type subsets for top feature types: populated places, streams, mountains, and hills. Our main findings are that wide discrepancies in coverage exist between the two datasets, in particular when country boundaries are crossed and when feature types with overall sparser coverage are concerned.

Contribution of the PhD candidate

I drafted the manuscript, wrote some of the R analysis code, wrote Python code for revisions, made all the maps and most of the figures. Stefano de Sabbata drafted part of the manuscript (mainly the 'Linear models' section), wrote some of the R analysis code, and obtained the original TGN (linked) data. Stefano de Sabbata and Ross S. Purves conceived of the original idea for the paper, which was further designed by the complete author team and implemented by Stefano and me.

Citation

Acheson, E., De Sabbata, S., and Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320. DOI:10.1016/j.compenvurbsys.2017.03.007

Computers, Environment and Urban Systems 64 (2017) 309–320

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A quantitative analysis of global gazetteers: Patterns of coverage for common feature types

Elise Acheson [a], Stefano De Sabbata [b], Ross S. Purves [a]

[a] University of Zurich, Department of Geography, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[b] Department of Geography, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom

ARTICLE INFO

Article history:
Received 21 September 2016
Received in revised form 28 January 2017
Accepted 13 March 2017
Available online xxxx

Keywords:
Gazetteers
Data quality
GeoNames
Placenames
Geocoding

ABSTRACT

Gazetteers are important tools used in a wide variety of workflows that depend on linking natural language text to geographical space. The spatial properties of these data sources, such as coverage, balance, and completeness, affect the performance of common tasks such as geoparsing and geocoding. However, little attention has focused on how these properties vary in global gazetteers, particularly across country boundaries and according to feature types. In this paper, we present a detailed investigation of the spatial properties of two open gazetteers with worldwide coverage: GeoNames, and the Getty Thesaurus of Geographic Names (TGN). Using point density maps, correlations, and linear regressions, we analyze the global spatial coverage of each data source for the full set of features and for top feature types: populated places, streams, mountains, and hills. Results show wide discrepancies in coverage between the two datasets, sharp changes in feature type coverage across country borders, and idiosyncratic patterns dominated by a few countries for the more sparsely covered natural features. As more and more systems rely on recognizing and grounding named places, these patterns can influence the analysis of growing amounts of online text content and reinforce or amplify existing inequalities.

1. Introduction

Gazetteers play a central role in linking text to space, influencing a multitude of application outcomes through their use in tasks such as identifying placenames[1] in text, disambiguating placename references, and associating placenames with a geographical footprint and type information. Until recently, gazetteers were primarily produced top-down, typically as curated resources for placenames in a prescribed area such as a country. Today, with data easily stored and shared online, and vast quantities of data released as open data, the ways in which gazetteers are being produced and distributed is evolving. At one end of the spectrum remains a top-down, strongly regulated process, where organizations such as national mapping agencies produce gazetteers according to explicitly defined data quality standards and local laws. At the other end are crowdsourcing efforts collecting information about places from anyone who wishes to contribute, often largely relying on the notion of the 'wisdom of the crowd' principles for data quality (Goodchild & Li, 2012). Somewhere on this spectrum are two gazetteers with some level of data curation, nominally global coverage, but limited explicit information with respect to data quality: GeoNames (GeoNames, 2016), and the Getty Thesaurus of Geographic Names (TGN, 2016).

These gazetteers form the focus of the present paper. Perhaps because of their worldwide coverage and their ready availability, both are popular in many research projects and applications, with GeoNames arguably the most commonly used gazetteer today. Despite this popularity, there has been limited scrutiny of its contents, with attention typically limited to a particular region or country, and focused largely on populated place features rather than a broader set of feature types. Smart et al. (2010) mapped the overall coverage of GeoNames in Great Britain, contrasting it with national mapping agency data and crowdsourced datasets. Ahlers (2013) conducted a broader examination of data quality in GeoNames, identifying anomalies and quality indicators for populated places in Central America, Germany, and Norway. Looking at both GeoNames and TGN, De Sabbata and Acheson (2016) quantitatively compared their coverage for all features and populated places in Great Britain, finding the datasets less detailed and less balanced than national mapping agency data. Although these studies have revealed that coverage in these products is unbalanced even within individual countries, the overall picture remains unclear since to date, an in-depth systematic global analysis, looking across country boundaries and at a range of feature types across gazetteers, has not been carried out.

An initial exploration of such properties examined global coverage of GeoNames alone and explored the distribution of a single feature type

E-mail addresses: elise.acheson@geo.uzh.ch (E. Acheson), s.desabbata@leicester.ac.uk (S. De Sabbata), ross.purves@geo.uzh.ch (R.S. Purves).
[1] We use the more vernacular term *placename* interchangeably with *toponym* in this paper.

http://dx.doi.org/10.1016/j.compenvurbsys.2017.03.007

(populated places) as a function of population (Graham & De Sabbata, 2015). Expanding on this work, we undertake a detailed comparative investigation of the global spatial properties of both GeoNames and TGN. We not only look at the full datasets, but also present a worldwide analysis of coverage for the four most frequent feature types in GeoNames, matched with corresponding types in TGN: populated places, streams, mountains, and hills. These four feature types account for a large portion of the full datasets in both gazetteers, particularly populated places which comprise over a third of all the data in both GeoNames and TGN. As for streams, mountains, and hills, they are among the most common natural features found in the data sets, and in the case of mountains, the most commonly referenced examples of a geographic feature in empirical experiments (Smith & Mark, 2001). Understanding the global coverage of these named natural features is particularly important in the context of any work analyzing the distribution of common toponym types (Campbell, 1991) and analysis of texts containing references to natural features (Moncla et al., 2014). For both gazetteers, we examine and compare feature distributions at fine, medium, and coarse granularities.

As discussed in the review that follows, *coverage* and *balance* are two pivotal quality indicators to assess the fitness for use of gazetteers for many common tasks. We therefore pose the following research questions:

1. How do GeoNames and TGN compare in terms of overall global coverage and balance?

2. How are important feature types in GeoNames and TGN distributed globally, and how do they compare in terms of coverage and balance?

We review previous work focusing on gazetteer properties, sources, and quality, as well as tasks in which gazetteers play a role. We then introduce in more detail the properties of the two gazetteers we analyzed, before setting out the analysis methods to characterize and compare GeoNames and TGN. Our results are presented as both graphical and numerical data, before we discuss their implications, particularly in terms of the suitability of these data sources for relevant tasks. We conclude with a list of key gazetteer shortcomings and propose future research focused on addressing these.

2. Gazetteers

> *"There is remarkable diversity in approaches to the description of geographic places (...)"*.
>
> [Linda Hill, Georeferencing, p. 94]

Gazetteers are resources that store structured information about places, minimally providing name, type, and location (or footprint) information for each place or record (Hill, 2000; Mostern et al., 2016). Each record may also contain other attributes such as alternative names, population information for populated places, and containment relationships - for example which country or region the place is in. Records may contain links to matching records in other datasets. These 'linked data' records are ones deemed to be about the same place through a matching process that, for instance, compares text, positional, and type information across resources (Sehgal et al., 2006; Smart et al., 2010). Placenames have in fact become a central node in linked open data, with GeoNames lying at the center of the linked open data cloud diagram (Schmachtenberg et al., 2014), demonstrating the efficacy of placenames as a way of relating information in the developing semantic web.

2.1. Gazetteer sources and production

Gazetteers have traditionally been produced in a top-down process, most commonly by national mapping agencies to serve as official placename resources for a defined area of interest such as a country, sometimes under specific legal or regulatory conditions. For example, Ordnance Survey (OS) produces the OS 1:50k gazetteer (2016) (and more recently, OS Open Names) for the extent of Great Britain, and SwissTopo produces SwissNames 3D (2016) for the extent of Switzerland. In the case of the United States, examples include a national resource for domestic names, the Geographic Names Information System (GNIS, 2016), developed by the U.S. Geological Survey, and an international resource for foreign names, the GEOnet Names Server (GNS, 2016), developed by the National Geospatial-Intelligence Agency.

As well as general purpose gazetteers, typically created by national mapping agencies and other government authorities, purpose-built gazetteers are created for a wide range of purposes. Among these are the TGN, a structured gazetteer with the aim of improving access to art, architecture, and material culture by enabling indexing. Due to its focus on these topics, historical names are important elements of the TGN, allowing links of historical artifacts to be made between present day locations and texts describing them in a historical context.

More recently, gazetteers have also been produced by incorporating bottom-up methodologies, where data is collected from multiple sources and integrated. Two heavily used global spatial datasets, OpenStreetMap and GeoNames, are produced this way: their sources include authoritative data, such as those described above where licensing permits, but also original data contributed by individuals, also known as volunteered geographic information (VGI) (Goodchild, 2007). Further still along the spectrum from top-down to bottom-up production are approaches to creating structured gazetteers using only crowdsourced data, through the extraction, analysis, and merging of multiple sources. One such example, the Gazetiki project, mined Wikipedia and Panoramio data to automatically create a gazetteer, relying on linguistic cues, search hits, and the GeoNames feature type hierarchy for entity typing (Popescu et al., 2008).

A complementary body of research focuses on both augmenting and enriching existing gazetteers and the generation of so-called meta-gazetteers to build better resources, whether more complete (with more features, or with richer annotation for existing features), or deemed more suitable for a particular task (Kessler et al., 2009; Smart et al., 2010). In one example using VGI, Gao et al. (2017) present a framework for efficiently creating new gazetteer entries from large numbers of user-tagged photographs, many of which contain feature types like 'park', 'museum', or 'river' as tags. Finally, OpenStreetMap has also been used as a gazetteer source directly, or to augment existing placename resources (de Oliveira et al., 2016; Hess et al., 2014; Yin et al., 2014).

As feature types are one of the three basic requirements of a gazetteer entry (Hill, 2000), any work seeking to integrate or augment gazetteers faces the challenge of assigning appropriate types to features, and potentially having to align different feature type ontologies to each other. A common use case in gazetteer conflation is to consider feature type information as evidence of (dis)similarity when trying to detect whether records are about the same feature (Fu et al., 2005; Hastings, 2008; Smart et al., 2010). However, this is a challenging task since feature types may vary widely between gazetteers, and the process of feature type alignment is itself complex (Janowicz & Keßler, 2008; Zhu et al., 2016). These difficulties are illustrated by for example Fu et al. (2005) who established "equivalence links" between feature type hierarchies, but found that strong constraints on feature type alignment led to poor performance. The underlying problem is further illustrated by Smart et al. (2010) who noted that even in national mapping agency data, large proportions of features were simply classified as "other". Zhu et al. (2016) recognize this challenge and combine top-down ontology analysis with bottom-up data-driven methods using spatial signatures related to instances of feature types to explore alignment issues in GeoNames, TGN and DBPedia Places.

2.2. Gazetteer quality

As introduced above, a wide variety of gazetteer and gazetteer-like resources exist, all of which may vary with respect to their data quality. In exploring gazetteer quality, we take as a starting point the so-called famous five, as listed by the US Federal Geographic Data Committee: attribute accuracy, positional accuracy, logical consistency, completeness, and lineage (Guptill & Morrison, 1995). Van Oort (2005) added to this list semantic accuracy, fitness for use, and temporal quality. In the more specific context of gazetteers, Leidner (2004) proposed seven criteria for gazetteer quality, a list extended and refined by Hill (2006, p.107) (Table 1).

An implicit quality which is not explicitly listed in the criteria above, but often mentioned in discussions of the nature of gazetteers, is coverage. In this paper we define the *coverage* of a resource as the feature density across space ('spatial coverage', as in Hill, 2006, p. 144). We define *balance* as in Table 1 as the uniformity of a resource across its scope of coverage, including the uniformity of its currency, accuracy, granularity, and richness of annotation. Thus balance and coverage are clearly related, since balance depends on the feature density across space (coverage) across the resource. As an example, a gazetteer covering features down to street-level detail in London but only down to neighborhood detail in Paris would be less balanced, but have better coverage in London, than a resource covering only neighborhood-level features in both cities. In the analysis that follows, we primarily focus on balance as the uniformity of coverage, as commenting on the uniformity of currency, accuracy, and richness of annotation is beyond the scope of this work.

An important upstream factor impacting gazetteer quality is the way in which the datasets are produced, as previously described. In the case of top-down datasets from mapping agencies, the organizations producing these resources typically ensure adherence to, and document, data quality standards, for example by sending surveyors out into the field in a structured manner, such that errors or omissions may be assumed to be randomly distributed in the dataset. However, this is not the case for crowdsourced data, which tend to show bias - that is, data quality which varies non-randomly as a function of the properties of the underlying space. For example, in a seminal paper on VGI quality, Haklay (2010) conducted a systematic analysis of OpenStreetMap data quality in terms of positional accuracy and completeness of street network data, and found geographical biases towards both urban (and therefore more populated) and more affluent regions in England, which have important implications for the balance of datasets produced through crowdsourcing in general. For gazetteers, such quality comparisons with authoritative datasets are possible for regions that have national mapping agency counterparts. However, a quality evaluation of global gazetteers must proceed in other ways, since no authoritative global database of placenames exists.

Table 1
Gazetteer quality criteria from Hill (2006).

Criterion	Description
Availability	"Degree to which the gazetteer is freely available and not limited by restrictive conditions of use"
Scope	"Small communal database, regional/national coverage, or worldwide coverage"
Completeness	"Degree to which the scope of the gazetteer is covered completely"
Currency	"Degree to which the gazetteer has incorporated changes"
Accuracy	"Number of detectable errors in names, footprints, and types"
Granularity	"Includes large, well-known features only or features of all sizes and those that are less well known"
Balance	"Uniform degree of detail, currency, accuracy, and granularity across scope of coverage"
Richness of annotation	"Amount and detail of descriptive information, beyond the basics of name, footprint, and type"

2.3. Using gazetteers

In general, gazetteers play a key role in tasks linking text to space, with applications ranging from disaster response or disease tracking through social media geolocation (Dredze et al., 2013; Zhang & Gelernter, 2014), to historical and literary text analysis (Cooper & Gregory, 2011; Southall et al., 2011). By enabling the organization of data according to geographical space, textual datasets can be spatially analyzed, opening up a wide range of possibilities for descriptive and predictive modelling of previously aggregated data.

More specifically, a number of distinct tasks require gazetteers or benefit from their use. A first task is the detection of placenames (or more broadly, geographic references[2]), for which a common approach is to look for textual matches between placenames in the gazetteer and each word, N-gram, or candidate placename in the text (Leidner & Lieberman, 2011). Thus, gazetteers and the placenames they contain influence whether a word or sequence of words is classified as a location in the first place (Leveling, 2015; Purves et al., 2007). Second, gazetteers are important in disambiguating placenames, as many disambiguation strategies use gazetteer data such as geometry, type, and attribute data (e.g. population) to rank candidates (Buscaldi, 2011). In fact, gazetteers typically determine whether a placename can be considered 'geo/geo' ambiguous, as this type of ambiguity in practice is defined as when a placename matches more than one entry in a gazetteer (Amitay et al., 2004; Zhang & Gelernter, 2014). Thus, toponym ambiguity is heavily influenced by the resources used, where gazetteers covering larger areas and finer granularities raise the potential for multiple matches (Buscaldi, 2011) - in other words, potentially increasing recall but decreasing precision.

The link between text and space is completed in a further task, focused on selecting a relevant geometry (footprint) for an input placename (Hill, 2000). This is crucial in geographical information retrieval, both to obtain geographical representations of textual queries and to index documents according to the geographical space they refer to (Purves et al., 2007). Of particular importance, setting aside geoparsing performance considerations, are the types of geometries available to model placenames. The simplest and most common geometry used to represent a placename is a point, but other geometries may be more appropriate according to place granularity and the nature of the reasoning to be carried out with the data (Alani et al., 2001; Guo et al., 2008). Currently, however, both GeoNames and TGN provide only latitude, longitude tuples as points referring to features in their free, downloadable versions.

An increasingly frequent text-to-space task is the geolocation of social media users or content. On Twitter for example, users may indicate a home location on their profile in free-text form. Thus, matching this text field to a gazetteer entry is often a key first step to analyzing Twitter users' attitudes and beliefs according to their location, using for instance sentiment analysis. Furthermore, since only approximately 1% of Twitter posts are explicitly tagged with GPS coordinates (Hecht et al., 2011), geocoding the posts themselves (or using the geocoded profile location as a proxy) can make more content available for analysis in use cases such as disaster response (Zhang & Gelernter, 2014) or disease tracking (Dredze et al., 2013). Jurgens et al. (2015) emphasize that gazetteer choice impacts the performance of geolocating Twitter users from the textual profile location information, with GeoNames returning matches for 500 k users and DBPedia only 75 k. Clearly coverage in general, and in particular balance with respect to the underlying properties of interest, are key indicators in assessing the quality of the results of such processes.

[2] Geographic references are considered a superset of placenames, which includes not only placenames but also place codes, such as addresses or postal codes, and more complex expressions, such as composite expressions like "North of Lake Ontario".

3. Data & methods

In order to obtain high recall for placename detection in text, it is important to use a gazetteer that provides both good coverage and completeness. To increase precision and limit the impact of ambiguity, it is furthermore fundamental to select a balanced gazetteer with an appropriate level of detail. This section first presents GeoNames and TGN, then describes the methods used to assess the coverage and balance of the two datasets.

3.1. GeoNames

GeoNames is arguably the most-used placename data source today, widely cited in academic works. It is not hard to understand why considering its unique combination of desirable properties: it contains over 10 million entries worldwide (coverage), is freely available online (availability), and has daily data exports (currency). Where its properties become less clear is concerning balance, data precision, completeness, and lineage. With respect to lineage, GeoNames consists of data originating from a variety of sources, some official sources and some individual contributors, but the source(s) for each particular record is not provided in the free version. This lineage issue muddles the already unclear picture with respect to balance, precision, and completeness. For instance, the true precision of a record may vary depending on the source or simply be unknown. As for balance and completeness, it is unclear to what extent each country or region is captured in the dataset, and thus the coverage may rather vary as a function of data availability for a particular area rather than as a function of the true concentration of named geographical features at those locations.

In addition to the standard elements of name, type, and geometry, GeoNames also provides for many entries a rich set of structured information including alternate names and spellings, population information for populated places, and hierarchical information such as containing administrative areas including countries. Geometries provided in the free gazetteer data download are latitude-longitude points for each feature, including for features with very large extents like countries and lakes. All features are classified according to a two-level type hierarchy consisting of a 9-feature-class top level and a 645-feature-code sublevel, with a short description provided for most feature codes. However, the distribution of feature counts is dominated by just a few of these 645 feature codes, with for example 3.4 million features having the code 'PPL' for 'populated place'. Though the feature type hierarchy has two levels, in practice a third-level is arguably encoded in the feature codes themselves, with for example 'STM' standing for 'stream', 'STMI' for 'intermittent stream', and 'STMIX' for 'section of intermittent stream'. We make use of this information when considering feature type alignments in Section 3.4.

3.2. TGN

TGN is a gazetteer resource developed for cultural heritage applications, freely available as linked open data since 2015. Unlike GeoNames, it is curated, has a stated focus on places of historical significance, and an intended use for "cataloguing, research, and discovery of art historical, archaeological, and other scholarly information" (TGN, 2015). Its coverage is nominally global and, appropriately for its historical focus, also covers a temporal range from "prehistory to the present".

With over 1.4 million entries of named places around the world, including over half a million populated places, TGN is a useful resource for text analysis, particularly for texts of a historical nature (Overell & Rüger, 2008; Smith & Mann, 2003). Similarly to GeoNames, TGN records provide names, feature type information, latitude-longitude coordinates, as well as hierarchical information where appropriate. Records also generally include sources and contributors, and may also contain descriptive notes and dates for historical places, as well as linkages between places signifying relationships such as 'successor of' and 'distinguished from'. TGN uses feature types from the Art & Architecture Thesaurus (AAT, 2017) type hierarchy, also from the Getty Research Institute. This type vocabulary and hierarchy features many more explicit levels than GeoNames and provides information about type semantics not only through a definition, but also with information about overlaps with other types and with lineage information about the type itself and its position in the hierarchy. In practice, TGN's feature type distribution is however also heavily skewed to a small number of types and again features a widely used category for population centers of all sizes, known as 'inhabited places'. In Fig. 1, the 'hills' type is shown in the AAT hierarchy with four sub-types, but while in our dataset 25,756 TGN features have 'hills' as a primary type, no TGN features have as primary type 'foothills', 'hillocks', 'hummocks (hills)', or 'knolls'.

3.3. Gazetteer quality criteria

In Table 2, we present a comparison of our two gazetteers of study and two representative national mapping agency gazetteers, OS 50k for Great Britain and SwissNames 3D for Switzerland, along nine quality criteria. These quality criteria are as in Table 1, with the addition of lineage (from the famous five of spatial data quality) and precision (which is documented explicitly for our national mapping agency data), and the removal of accuracy, which in practice is not available because it requires testing against a reference dataset. Though coverage does not appear explicitly in the table, our definition relates it to scope, completeness, granularity, and balance as discussed above.

These four datasets all share the characteristic of being freely available (though licensing conditions vary), but in the other dimensions, TGN and GeoNames are more similar to each other than either is to the national mapping agency datasets. Importantly, completeness and balance are unknown for GeoNames and TGN, not being explicit aims for either data source. Precision of the feature coordinates is not documented, though TGN states that their coordinates are approximate only. On the other hand, GeoNames and TGN offer the advantage of nominally rapid update cycles, with GeoNames providing daily data downloads online. The authoritative datasets have slower release cycles, consistent with a process requiring extensive quality control to ensure completeness, precision, and balance of their contents. Our analysis of GeoNames and TGN aims to enable informed statements about balance and coverage.

3.4. Feature type selection and matching

Full data snapshots of both GeoNames and TGN were obtained on June 30th 2015. Though differences in feature density are to be expected, since GeoNames has almost 10 times the number of features as TGN,

Fig. 1. AAT type hierarchy containing the feature type 'hills', used in TGN. (Source: AAT, 2017).

Table 2
Gazetteer quality criteria for four gazetteers.

Criterion	GeoNames	TGN	OS 50 k	SwissNames 3D
Availability	Free	Free	Free	Free
Scope	Worldwide	Worldwide	Great Britain	Switzerland
Completeness	?	?	✓	✓
Currency	Daily	Two weeks	Annual	Annual
Precision	Varied	Approximate	1k grid cell	0.2 m–3 m
Granularity	Medium to fine	Medium	Medium to fine	Fine
Balance	?	?	Uniform	Uniform
Richness of annotation	Medium	Rich for portion	Medium	Medium
Lineage	Various sources	GNIS, experts	OS maps	SwissTopo maps

we expect datasets to become more similar as they tend to more accurately sample the real world, and in the future, as they become more integrated through the practice of linked data. Our snapshot of TGN was taken shortly after the product was launched in its linked open data form, and before it had had time to be propagated into GeoNames.

A first step towards our goal to compare GeoNames and TGN was to match countries from one dataset to the other, which was done manually and resulted in a set of 237 common countries which could be used to select features by country from the raw gazetteer data. A second step was to select feature types for analysis, a process primarily driven by looking at the most common features types in GeoNames and matching these to types in TGN.

Aligning feature types between gazetteers is a complex problem, as resources may have different feature type ontologies, the same words may take different meanings across resources, and the meaning of a particular word may also itself vary among geographical regions and individuals (Zhu et al., 2016). In our alignment process we considered type names, definitions, feature type hierarchies and, since our analysis of coverage and balance relies on feature counts, type frequencies in each dataset. Given the potential influence of feature alignment on any results comparing gazetteer content, we furthermore carried out a series of sensitivity tests to explore such effects, where we varied the alignment choices using one-to-one, one-to-many, many-to-one, and many-to-many links between GeoNames and TGN types.

Fig. 2 shows the most frequent feature types in each gazetteer (using unique feature codes for GeoNames), the results of our feature type selection and alignment, and the number of features of each type selected for analysis.

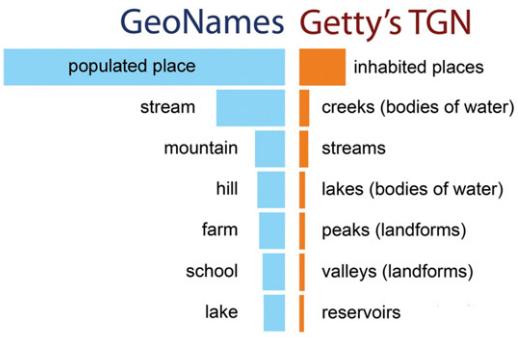
Based on our sensitivity tests, we aligned 'populated place' (PPL) and its implicit sub-types (PPL*) in GeoNames to the 'inhabited places' place type in TGN, and proceeded similarly for the other types, as shown in Fig. 2. Whereas GeoNames has a large number of features typed with implicit sub-types (for example, PPLL is used for 141,798 features and STMI for 177,531 features), TGN has very low counts for sub-types, perhaps because of its richer hierarchical type structure and its use of a preferred place type for each feature with optional secondary types.

3.5. Analysis methods

For a multi-scale, feature-type specific understanding and comparison of the global coverage of features in GeoNames and TGN, three analysis scales were chosen. At the finest level, 10x10km cells with 30 km neighborhoods were chosen as the point aggregation unit for global point density maps. For the full set of features ('all') and for the four selected feature types ('populated places', 'mountains', 'hills', 'streams'), maps were produced using the ArcGIS Point Density tool, in the Goode Homolosine Land equal area projection, resulting in 10 global maps. These maps allow for a visual overview of the global coverage of features in each of GeoNames and TGN, and a visual comparison of this coverage between the two gazetteers for each feature type. Furthermore, by exploring differences in coverage, it is possible to gain some insight into balance (for example where feature density varies greatly in a dataset across national boundaries).

For the second part of the analysis, features were aggregated at two coarser-grained spatial units: 100 × 100 km cells, and individual countries. Aggregated counts for both spatial units were calculated for each feature type (five in total, including 'all'). The country of each feature was available directly as attribute data in both datasets, thus country counts could be obtained by summing the number of features with each country attribute. For the 100 × 100 km cells, counts of features in each cell were obtained through a spatial join. For each aggregation unit (2) and each feature type (5), counts in the two datasets were plotted and correlation coefficients computed using ranks (Kendall's method), since values were not normally distributed.

Based on the outcomes of the second part of the analysis, linear models relating the two gazetteers were calculated to establish descriptive, quantitative relationships between their coverage. Through linear models, feature coverage can be compared not just on the basis of



Feature type name	GeoNames feature codes	TGN place types	GeoNames count	TGN count	GeoNames % of total	TGN % of total
all	(all)	(all)	10,203,772	1,439,138	100.00%	100.00%
populated places	PPL*	'inhabited places'	3,767,721	564,533	36.92%	39.23%
streams	STM*	'streams'	1,027,913	108,445	10.07%	7.54%
mountains	MT*	'mountains'	390,418	25,975	3.83%	1.80%
hills	HLL*	'hills'	351,926	25,680	3.45%	1.78%

Fig. 2. Most frequent feature types for GeoNames and TGN (top); Types selected for analysis and their respective counts and frequencies in GeoNames and TGN (bottom). (Only features considered for the country and graticule analyses are included in the table. Excluded features, not in any of the matched countries, accounted for 0.26% of all features GeoNames and 0.15% in TGN.) * Indicates that all GeoNames feature codes taking this base were included.

rank, but also magnitude, by describing a proportional relationship between feature counts in corresponding cells or countries.

4. Results and interpretation

4.1. Point density maps

We first present the global point density maps of all the features in each of GeoNames and TGN (Fig. 3). These maps show the density of placenames in each data source using quantiles calculated on the GeoNames data for both maps to ease comparison. Overall, for both data sources, we observe higher densities of placenames in regions such as Europe and in the eastern United States, and much lower densities in entire continents such as South America and Africa. These global differences in coverage are particularly marked in the TGN data, where there is widespread data scarcity in South America with the exception of Chile, and likewise for Africa aside from a concentration of placenames in Egypt, a place of great historical significance. Another observation common to both maps, but particularly pronounced in TGN, is that in many places coverage seems particularly uneven across country borders. For example, GeoNames placename density is markedly different between Norway and Sweden, and in TGN the eastern and southern borders of Germany are clearly distinguishable due to a sudden drop in coverage, and similarly for the border between Canada and the United States. India is described in relatively similar detail by GeoNames and TGN, whereas the amount of features in neighboring Nepal and Sri Lanka is strikingly different in the two datasets. These results also indicate that balance is an issue in both datasets, since these patterns are unlikely to reflect real toponym density.

Breaking down the datasets by feature type helps shed light on whether these observations are consistent across, or driven by, particular types. Fig. 4 shows small multiples of coverage in GeoNames and TGN for each of the four feature type data subsets, starting at the top with the most frequent feature type, populated places, down to the least frequent (in GeoNames), hills. This sequence illustrates that as a feature type decreases in numbers overall in a gazetteer, its global coverage also becomes sparser, concentrated in a smaller area of the globe. Whereas in GeoNames populated places show non-zero density across large swaths of the globe, most 10 km density cells are zero (white) for mountains and hills. This observation is more pronounced in TGN,

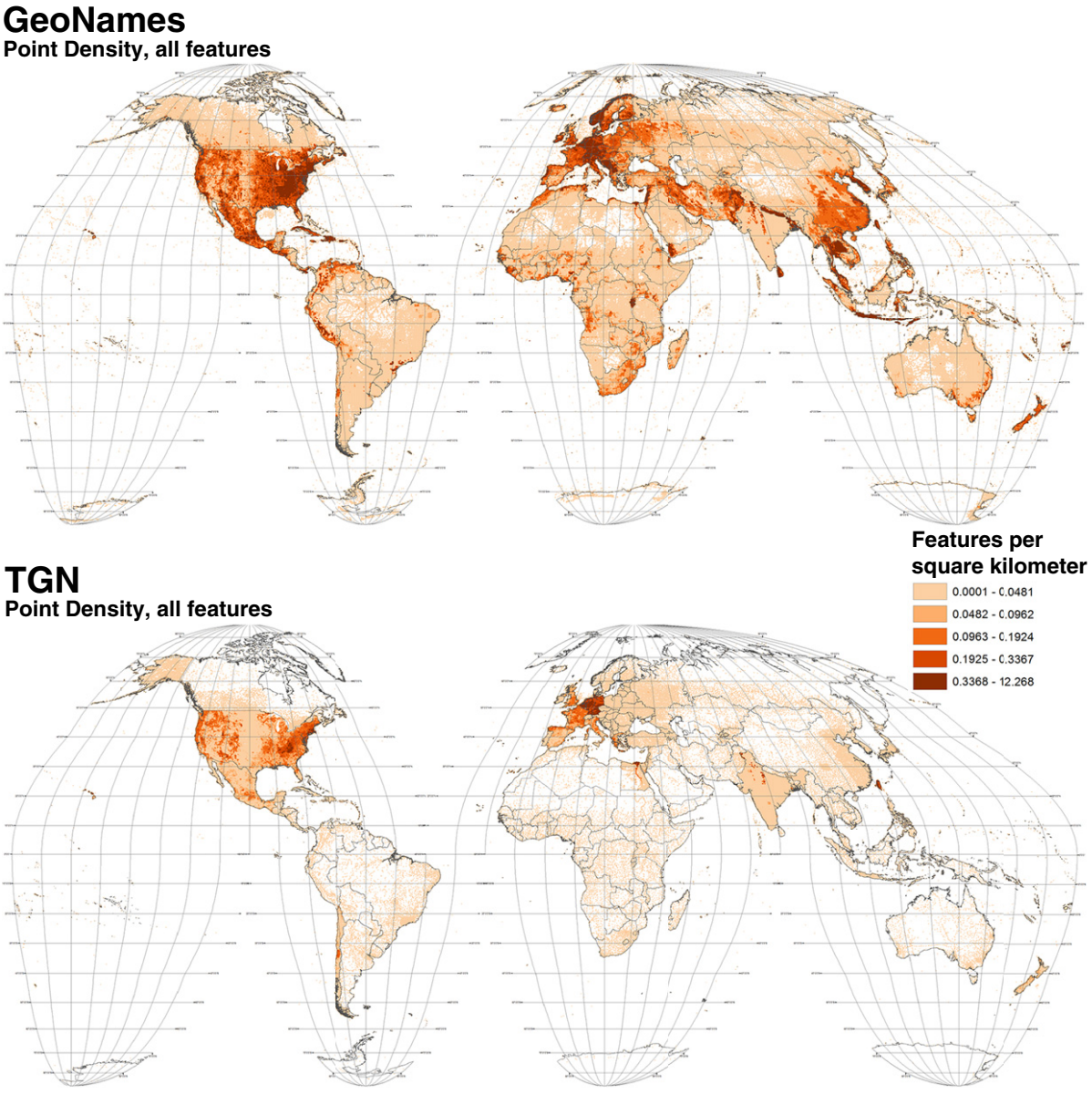


Fig. 3. Point density maps for all features in GeoNames (top) and TGN (bottom) rendered in terms of GeoNames quantiles, in the Goode Homolosine Land projection.

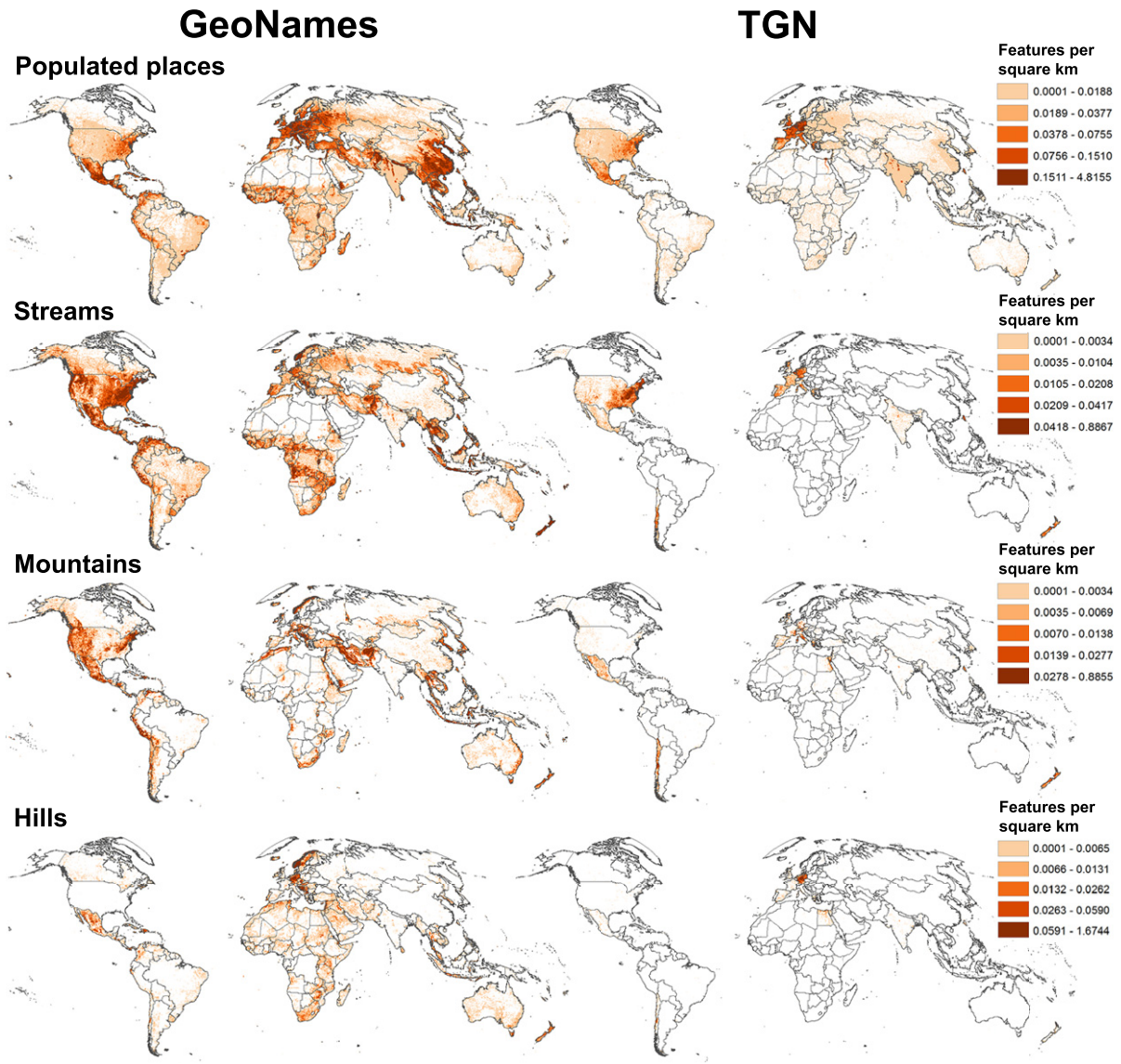


Fig. 4. Point density maps by gazetteer (GeoNames, TGN) and feature type (populated places, streams, mountains, hills), rendered in terms of GeoNames quantiles, Goode Homolosine Land projection.

which overall has about ten times less data than GeoNames in terms of all features, but is also slightly more skewed towards populated places than GeoNames with 39% of its features populated places but only 1.8% hills, compared to 37% and 3.4%, respectively, for GeoNames (Fig. 2). While the TGN populated places map again clearly shows the resource's emphasis on Europe and the United States, the global TGN hills map shows just how few features of this type are catalogued.

In order to address whether the coverage of features in GeoNames and TGN corresponds well to the true distribution of named features in the world, it is helpful to consider feature types in isolation. For the best-represented feature type, populated places, population density may be a reasonable proxy for density of named populated places at our scales of analysis (Graham & De Sabbata, 2015). In both the GeoNames and TGN populated places maps, some areas of visibly high feature concentration correspond with areas of high population density such as continental Europe and the North East of the United States, but other populous regions seem comparatively less well-covered, including India, Brazil, and the African continent in general. TGN in particular appears to have a strong focus on Western Europe and the United States in terms of catalogued populated places, and relative data poverty through China, South-East Asia, and parts of the Middle East compared to GeoNames.

Shifting down from the best-represented feature type to the sparsest type in our analysis, hills, the uneven coverage becomes more extreme, with virtually no hills catalogued in the United States (both in GeoNames and TGN), and a very high concentration of hills in Norway (GeoNames) and Germany (GeoNames and TGN). These maps again show sharp changes in feature type coverage at country borders, such as the US-Mexico border for hills (GeoNames) and mountains (TGN), and the US-Canada border for streams (GeoNames). A closer look at the individual point features shows a dearth of mountains in TGN for Switzerland (Fig. 5b), a country renowned for its mountains, and virtually no hills in GeoNames for the US (Fig. 5c) against an abundance in neighboring Mexico. Thus it appears from these maps that the country unit is an important driver of global coverage by feature type.

To explore this behavior where these feature types were seemingly well sampled, counts by type were plotted for the ten countries with the highest numbers of the given types (Fig. 5a). These bar charts in all cases show that one or a few countries dominate the distribution for a particular feature type, for instance Norway for hills in GeoNames

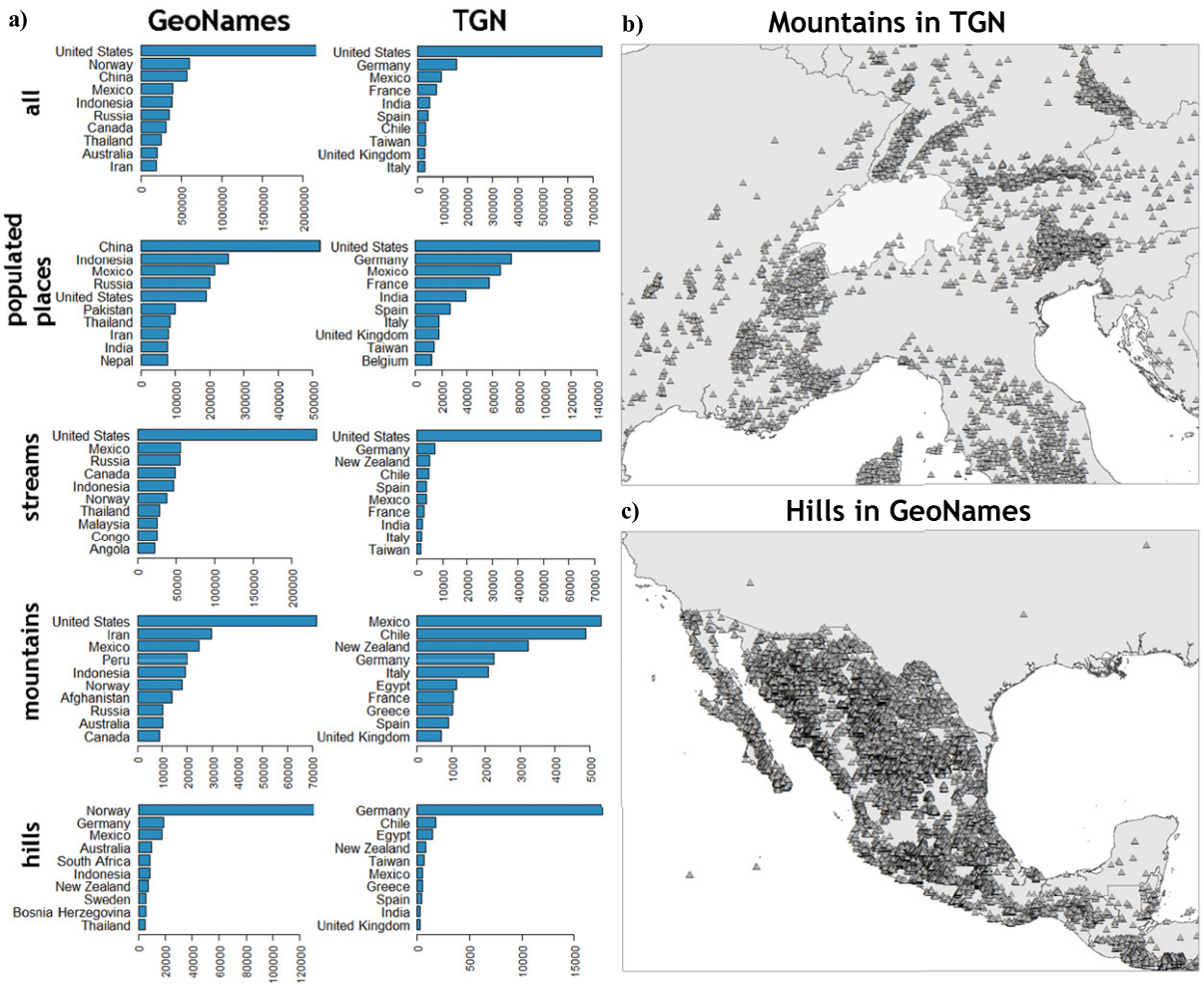


Fig. 5. a) Bar charts of the ten countries with the most features of each type; b) all mountain features in TGN for Switzerland (white) and surrounding countries; c) all hill features in GeoNames for Mexico and the southern United States.

or Germany for hills in TGN. Furthermore, GeoNames and TGN tell different stories about the underlying distribution of data, especially for the natural feature types. Indeed, Norway unequivocally tops the list for hills in GeoNames, yet does not even make the top ten in TGN. Similarly, China is conspicuously absent from the top ten list for populated place features in TGN, but tops the list for populated places in GeoNames. All of these observations again emphasize the lack of balance for all the analyzed feature types in both GeoNames and TGN.

4.2. Correlations between GeoNames and TGN

To compare coverage between GeoNames and TGN systematically, counts of features were spatially aggregated according to the coarser, meaningful unit of countries, and the finer unit of 100 × 100 km cells (created in the equal area Goode Homolosine Land projection). Corresponding counts were analyzed using Kendall's tau rank correlation, with the results shown in Table 3.

Table 3
Kendall's tau correlation coefficients between GeoNames and TGN for countries and raster cells.

Features	Countries (N = 237)		Raster (N = 51,996)	
	M (neither 0)	Kendall's tau	M (not both 0)	Kendall's tau
All	237	0.7138*	20,665	0.6383*
Populated places	235	0.7015*	14,188	0.5377*
Streams	29	0.3004+	13,182	0.2322*
Mountains	159	0.4853*	9868	0.2449*
Hills	74	0.2584^	8704	0.0424*

* $p < 0.00001$.
^ $p < 0.01$.
\+ $p < 0.05$.

For the country counts, a pair was included in the rank correlation calculation when neither country had a count of zero to avoid having artificially high correlation coefficients due to matching pairs with very low or zero counts. For the much larger number of raster cells, a pair was included in the rank correlation calculation when either dataset had a non-zero count, striking a balance between keeping spurious pairs with no data and dropping meaningful pairs.

The correlation coefficients for feature counts by country show relatively strong positive relationships when accounting for both all features and populated places, and a weaker but still positive relationship for mountains. For streams and hills the relationships were much weaker. The correlation coefficients for raster cell counts show similar patterns, with the highest correlations for all features and populated places, and much weaker or no relationships for streams, mountains, and hills. Comparing correlation coefficients for countries and raster cells, we note that the coefficients are greater for countries than raster cells in all five cases, meaning a stronger relationship exists at the country level than for the finer raster cells.

4.2.1. Sensitivity to feature type alignment

In order to ensure our results were robust to changes in feature type alignment, we performed sensitivity tests where we selected different combinations of feature types from GeoNames and/or TGN and

repeated the correlation analysis. We varied GeoNames feature subsets by selecting only a single dominant feature code (PPL, STM, MT) for a type rather than also including its implicit sub-types (PPL*, STM*, MT*). As for TGN, of particular interest was the inclusion of types featured in relatively high numbers in the collection: creeks and peaks.

Table 4 illustrates the results of these sensitivity tests. In all but one case (where streams + creeks from TGN are included) correlations are similar and statistically significant. In the case of creeks, 99.6% of all creeks in the TGN are located in the USA. These results suggest that our choice of feature alignment is robust.

4.3. Linear models

The Kendall rank correlation coefficients indicated the existence of varying degrees of positive relationships between aggregated features counts in GeoNames and TGN, depending on type and unit of analysis (country or raster cells). In order to analyze these relationships considering not only the rank, but also the magnitude of feature counts, linear regression models were used.

Based on the positively skewed distribution for both GeoNames and TGN aggregated counts, we used log-log regression models, arbitrarily using GeoNames counts as the independent variable and TGN counts as the dependent variable. While raw counts suggest that TGN contains only 14% of the amount of features in GeoNames, the first two stages of the analysis suggest that the relationship is more complex. The coefficients (b) of the regression models provide a better estimate of the quantitative relationship, whereas the coefficient of determination (R^2) provides a measure of model fitness.

As such, the linear models presented in this section should not be interpreted as explanatory or predictive, but rather as descriptive. These two data sources are clearly independently produced, but as they are both sampling geographical features from the real world, relating their feature counts can give us insight into how similar their coverage is. The selection of one variable as dependent and the other as independent is purely arbitrary, and does not affect the interpretation of the results.

A first linear model was constructed, using countries as the unit of analysis, based only on the independent and dependent variables mentioned above, but the model did not meet the assumption of homoscedasticity of the residuals. This was confirmed through a Breusch-Pagan test, which was significant ($p < 0.001$). The log-log scatter plot of TGN vs GeoNames for all features showed an interesting pattern where a group of countries showed relatively high counts in TGN compared to the remaining countries, as depicted in Fig. 6. From this scatter plot, we identified this set of 15 countries as 'high coverage' countries: United States, Germany, Mexico, France, India, Spain, Chile, Taiwan, United Kingdom, Italy, Egypt, Greece, Belgium, New Zealand, and the Netherlands. We found that these same countries were also well covered across the four feature types we analyzed. The data point for the Faroe Islands was excluded as a statistical outlier. A dummy variable (i.e., Boolean indicator) was thus introduced, taking the value 1 when the country is a member of the so-called *HighCoverage* set, and 0 otherwise. We then used linear models of the form:

$$\ln(\mathrm{TGN}) = b_0 + b_1 \ \ln(\mathrm{GeoNames}) + b_2\ \mathrm{HighCoverage} + \varepsilon$$

The linear model then obtained is presented in the first section of Table 5 (Model 1).

Model 1 in Table 5 is robust and fit. The residuals are normally distributed (Shapiro-Wilk test, $W = 0.99$, $p > .01$), satisfy the homoscedasticity assumption (Breusch-Pagan test, $BP = 4.27$, $p > .05$), and the errors are independent (Durbin-Watson test, $DW = 1.88$, $p > .05$). No statistically influential cases were identified. The number of features in GeoNames, combined with the distinction between high coverage and low coverage countries, accounts for 87% ($F(2,232) = 803.6$, $p < .001$) of the variation in the number of features in TGN, when aggregated by country.

This model illustrates how the number of features in TGN in high coverage countries is of the same order of magnitude as counts in GeoNames, having over 60 ($e^{4.13} = 62.18$) as much content as low coverage countries. Still even in the high coverage group, an increase of 100 features in GeoNames corresponds to an increase of only 71 features in TGN. The model also supports the two assumptions discussed in this section. First, a clear relationship exists between the two datasets. Second, two distinguishable groups of countries are present in TGN, one featuring an amount of content comparable with what can be found in GeoNames, and another group covered in far less detail. The characteristics of the high coverage group are further discussed in the next section.

Based on these results, we investigated the possibility that linear models might hold at a different scale of analysis. We again used the 100×100 km raster cells that we used in the correlation section and relate the number of features in each cell for GeoNames and TGN. We tested a linear model similar to the model presented above, assigning each cell to a country (or no country where necessary) and re-creating the dummy variable *HighCoverage* as above. All cells with a count of zero for either gazetteer were discarded, due to the logarithmic nature of the models. Given the large number of cells remaining (13,910) we also tested the same model with smaller random samples (150 and 1500 cells), which resulted in consistent outcomes and significance levels.

The linear model based on the raster cell counts (Model 2 in Table 5) is very similar to the model based on country counts presented above (Model 1), when disregarding cells not associated with any country. The number of features in GeoNames (combined with the dummy variable) accounts for 82% ($F(2, 11{,}152) = 0.00026$, $p < 0.001$) of the variation in the number of features in TGN. Similarly to the model above, high coverage cells contain about thirty ($e^{3.40} = 29.96$) times as much content as low coverage cells in TGN. The residuals are normally distributed, but they show heteroscedasticity, and errors are not independent. This is most probably due again to a different behavior between countries in the high coverage and low coverage sets. Including the cells which are not associated to any country (Model 3 in Table 5) leads to a lower $R^2 = 0.75$ ($F(3, 13{,}906) = 0.00014$, $p < 0.001$), while the remaining values are stable.

Finally, we tested similar linear models based on number of features per country, for the feature types populated places and mountains. These linear models showed a relatively strong relationship between the two gazetteers, but the residuals of the models were not normally distributed. It is also important to note that these models showed substantial linear relationships only when excluding countries with no features in at least one gazetteer, as TGN in particular contains no features of those two types for many countries (counts of zero are found in 34 countries for populated places, 99 countries for mountains, 172 countries for hills, and 209 countries for streams).

Table 4
Kendall's tau correlation coefficients between GeoNames and TGN for countries using different feature type alignments. Alignments used in the rest of the paper are highlighted.

Feature type name	GeoNames feature codes	TGN place types	M (neither 0)	Kendall's tau
Populated places	PPL*	Inhabited places	235	0.7015^{α}
Populated places	PPL only	Inhabited places	232	0.6996^{α}
Streams	STM*	Streams	29	0.3004^{+}
Streams	STM only	Streams	29	0.3265^{+}
Streams	STM*	Streams + creeks	42	0.1757
Mountains	MT*	Mountains	159	0.4853^{α}
Mountains	MT only	Mountains	159	0.4763^{α}
Mountains	MT*	Mountains + peaks	164	0.4946^{α}

$^{\alpha}p < 0.00001$, $^{\wedge}p < 0.01$, $^{+}p < 0.05$.

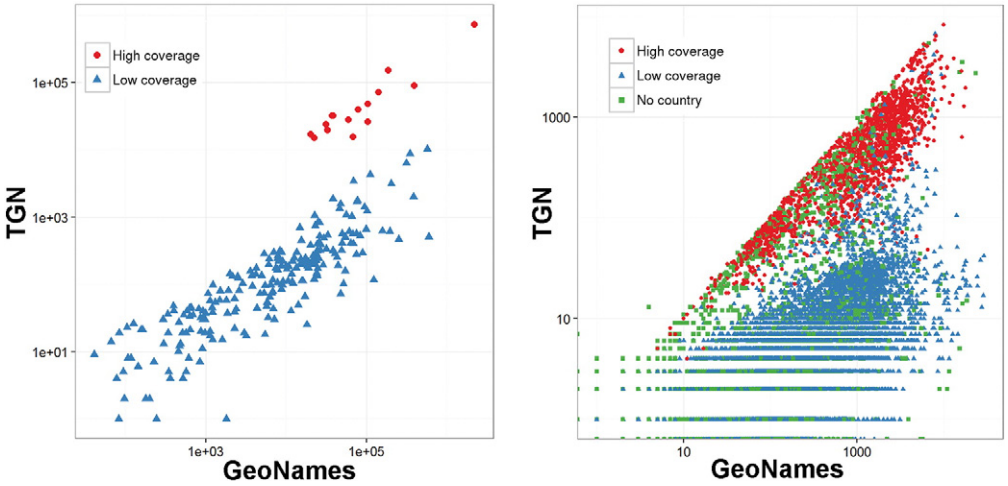


Fig. 6. Log-log scatter plot of feature counts in TGN as a function of counts in GeoNames in matching countries (left) and 100 × 100 km cells (right).

4.4. High coverage countries

The linear models from the previous section showed a systematic pattern where a group of countries were consistently better represented in TGN than the others. We termed this group of countries 'high coverage' and computed the Kendall's tau correlation coefficients for this list of 15 countries in isolation, as we did for all countries in Section 4.2. The resulting values, presented in Table 6, show strong highly significant positive relationships between these high coverage countries in TGN and GeoNames. This is the case not only when considering all features, which was expected based on the models presented in the previous section, but also for the four feature types taken into account.

The correlation coefficients are not only higher in each case for only the list of 15 as compared to the full set of countries with non-zero feature counts, but the values are also much less variable, all ranging from about 0.70 to 0.87. Overall these correlation coefficients show that not only do these high coverage countries have more data in TGN than the rest of the countries, with few exceptions, but also that where TGN coverage is high, the data resembles GeoNames much more, in terms of the feature counts per country.

Table 5
Linear models of feature counts in TGN as a function of feature counts in GeoNames.

Models	Adj. R^2	Coefficients (b)	Std. Error	Std. coef. (β)	p (sign.)
1) Country counts					
TGN	0.87				
Constant		−1.48	0.226		<0.001
GeoNames		0.71	0.025	0.68	<0.001
High coverage (dummy)		4.13	0.218		<0.001
2) Raster counts (countries only)					
TGN	0.82				
Constant		−1.22	0.027		<0.001
GeoNames		0.51	0.005	0.43	<0.001
High coverage (dummy)		3.40	0.022		<0.001
3) Raster counts (all non-zero)					
TGN	0.75				
Constant		−1.26	0.027		<0.001
GeoNames		0.51	0.005	0.48	<0.001
High coverage (dummy)		3.59	0.026		<0.001
No country (dummy)		0.67	0.022		<0.001

5. Discussion

"Are my data fit for purpose?" This is a crucial question in science, and in this paper we set out to illustrate how it applies to the use of two global gazetteers. Coverage and completeness play a fundamental role in the detection of placenames in text, particularly in terms of recall, while balance is important in limiting ambiguity and improving precision. The maps presented in the first part of the previous section illustrate the skewness and idiosyncrasies of two major global gazetteers, GeoNames and TGN. The correlation and regression analysis presented above shows how the two gazetteers do not provide a coherent description of the world's toponyms, and identify regions and scales where similarities do exist.

As TGN is a historically-focused, curated resource possessing about a tenth of the overall quantity of features in GeoNames, differences in placename density must exist. Our methods allowed us to identify a small list of countries whose placenames are catalogued in more detail in TGN than the others - but these still only possess a portion of the quantity of data provided by GeoNames. Among those countries is the United Kingdom, where TGN provides only half the placenames available in GeoNames, which in turn are just a fraction of the data provided by the Ordnance Survey (De Sabbata & Acheson, 2016). Beyond such countries with detailed coverage, the number of features available in TGN drops to two orders of magnitude lower than in GeoNames, consistent with the results reported by Ahlers (2013) for the specific case of Honduras. Thus, TGN coverage is not only sparser overall, but more idiosyncratic than GeoNames and thus also less balanced.

Both GeoNames and TGN differ fundamentally from gazetteers produced by mapping agencies, which adhere to defined data quality standards including completeness and balance, and whose contents can be assumed to vary largely as a function of the true density of named features in the area of interest. Indeed, one of the most challenging issues, which we can only address peripherally in this paper when exploring global gazetteers, is completeness. Our results suggest that GeoNames and TGN are both far from complete given variation in coverage and feature type balance, and based on our results we can suggest regions

Table 6
Kendall's tau correlation coefficients for high coverage countries in TGN.

Features	N	Kendall's tau
All	15	0.6952*
Populated places	15	0.8667*
Streams	15	0.8476*
Mountains	15	0.7905*
Hills	15	0.7656*

* $p < 0.001$.

where the datasets may be particularly incomplete. However, understanding whether toponyms are missing because they have not been mapped, or are simply not used, requires us to also consider both the underlying physical landscape and variation in toponym usage across cultures and languages (Burenhult & Levinson, 2008). As for balance, GeoNames coverage varies partly as a function of the availability of data, with rapid changes in coverage possible overnight when new datasets are integrated, thus affecting the balance of the resource. TGN represents a historically-focused view of the world, but even then some coverage artifacts seem tied to open data integration such as the relative abundance of data in New Zealand and of hills in Germany. Future work studying the lineage of the features could explain some of these observations and perhaps reveal crucial information on common sources between the gazetteers.

Indeed, our results clearly highlight the role of institutional - usually national - open data, as the coverage offered for feature types shows abrupt changes across national borders. Throughout the analysis, countries consistently appeared as the strongest driver of variation. Even at the finest analysis scale, the influence of the country unit was visible in the global maps for both GeoNames and TGN. Therefore, studies assessing the quality of gazetteers through comparison with authoritative datasets cannot simply be generalized to other study areas, particularly across borders. Furthermore, special care should be taken when working in a multi-national study area (for example Europe), as taking gazetteer data as-is across borders will typically result in variations in balance with respect to the sampling of the true distribution of named places, and results of tasks such as geoparsing are more likely to reflect gazetteer properties, rather than true spatial variation. In any work seeking to augment gazetteers, combine gazetteers, or create meta-gazetteers (Gao et al., 2017; Grossner et al., 2016; Smart et al., 2010), an important aim should be not to introduce further bias in coverage or balance.

Another important observation is that coverage and correlations between feature types quickly decreases for all except populated places, which account for over a third of all data in GeoNames or TGN. Our maps show that as overall numbers in one gazetteer for a particular feature type decrease, coverage across the globe becomes not only sparser, but less well correlated, more idiosyncratic and thus less balanced, even for the common natural feature types streams and hills. An analogy can be made with crowd-sourced mapping projects, which have been shown to suffer from biases that are not only geographic, but also thematic. Bégin et al. (2013) notes that natural features tend to suffer from lower positional accuracy than man-made features in VGI, and finds that users in mapping tasks show preferences for mapping certain feature types. Similarly, our results show that the representation of natural features is of a comparatively lower quality than populated places in GeoNames and TGN, the data sources being more focused on populated places globally. Thus, we suggest national data should be used preferentially when dealing with these feature types, particularly since national borders in any case are a strong driver of coverage, removing any advantage of nominally seamless, global, datasets. Furthermore, though our sensitivity tests indicate that our feature type alignment choices for very common features are robust in calculating correlations, the importance of alignment (Zhu et al., 2016) in matching less frequent features types is likely to have a bigger influence on gazetteer quality assessment.

Though our results clearly indicate that balance for natural feature types is poor, quantifying this further requires the use of meaningful proxies for named features. Possible approaches, which might also give insights into completeness, might imply modelling expected densities of named features as a function of morphological properties (Hengl & Reuter, 2008) and relating this to existing authoritative gazetteer data. Finally, current research in information geographies suggests that most datasets are heavily skewed towards the Global North and marginalize the Global South (Graham et al., 2015) – including those used to train machine learning algorithms – rendering any aim for a global, rich, balanced gazetteer a formidable challenge.

6. Conclusions

The present paper has illustrated the important role played by gazetteers in the current data revolution, and argues that fitness for use of global gazetteers has been neglected, despite their very common application to a wide range of tasks. Our results highlight the skewness and idiosyncrasies of these gazetteers whose coverage and balance, especially at the level of feature types, varies widely, and is best predicted by national borders. These results also suggest that the politics and economics of open data (Kitchin, 2014) can have a significant impact on gazetteers, and thus on geographic analysis and automated data processing.

Although making an informed decision on fitness for purpose is straightforward with top-down, authoritative gazetteers through the documented quality criteria, this is in practice much more difficult with global gazetteers such as GeoNames and TGN. In such a situation, the questions that researchers using the gazetteers discussed in this paper should ask themselves are: what components of a pattern extracted by linking text to space reflect meaningful patterns in the underlying data, and what simply reflects skewness and idiosyncrasies of the gazetteer used in the analysis?

Haklay's (2010) conclusion that "places where population is scarce or deprived are, potentially, further marginalised by VGI exactly because of the cacophony created by places which are covered" seems to merit an extension to the realm of global gazetteers. Less connected and less developed countries are currently further marginalized in global gazetteers, as are natural features, drowned out by populated places. The cacophony of information is further intensified by algorithmic data analysis, which through the use of gazetteers produce even more data about the places already covered.

Acknowledgements

This work was supported by an STSM Grant from COST Action IC1203.

References

AAT (2017). Art & architecture thesaurus online. Retrieved from http://www.getty.edu/research/tools/vocabularies/aat/ (Accessed 25.01.2017)

Ahlers, D. (2013). Assessment of the accuracy of GeoNames gazetteer data. *Proceedings of the 7th Workshop on Geographic Information Retrieval* (pp. 74–81). New York, NY, USA: ACM GIR '13 10.1145/2533888.2533938

Alani, H., Jones, C. B., & Tudhope, D. (2001). Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science*, *15*(4), 287–306. http://dx.doi.org/10.1080/13658810110038942.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging web content. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 273–280). New York, NY, USA: ACM SIGIR '04 10.1145/1008992.1009040

Bégin, D., Devillers, R., & Roche, S. (2013). Assessing Volunteered Geographic Information (VGI) quality based on contributors' mapping behaviours. *Proceedings of the 8th International Symposium on Spatial Data Quality ISSDQ* (pp. 149–154).

Burenhult, N., & Levinson, S. C. (2008). Language and landscape: A cross-linguistic perspective. *Language Sciences*, *30*(2–3), 135–150. http://dx.doi.org/10.1016/j.langsci.2006.12.028.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, *3*(2), 16–19. http://dx.doi.org/10.1145/2047296.2047300.

Campbell, J. C. (1991). Stream generic terms as indicators of historical settlement patterns. *Names*, *39*(4), 333–366. http://dx.doi.org/10.1179/nam.1991.39.4.333.

Cooper, D., & Gregory, I. N. (2011). Mapping the English Lake District: A literary GIS. *Transactions of the Institute of British Geographers*, *36*(1), 89–108. http://dx.doi.org/10.1111/j.1475-5661.2010.00405.x.

De Sabbata, S., & Acheson, E. (2016). Geographies of gazetteers in Great Britain. *Proceedings of the 24th GIS Research UK Conference, GISRUK 2016. Greenwich, UK.*

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)* (pp. 20–24).

Fu, G., Jones, C. B., & Abdelmoty, A. I. (2005). Building a geographical ontology for intelligent spatial search on the web. *Proceedings of IASTED international Conference on Databases and Applications (DBA-2005)* (pp. 167–172). Innsbruck, Austria: ACTA Press.

Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered big geo-data based on hadoop. *Computers, Environment and Urban Systems, Geospatial Cloud Computing and Big Data*, *61*(Part B (January)), 172–186. http://dx.doi.org/10.1016/j.compenvurbsys.2014.02.004.

GeoNames (2016). Retrieved from http://www.geonames.org (Accessed 05.10.2016)

GNIS (2016). United States Board on Geographic Names: Domestic names. http://geonames.usgs.gov/domestic/index.html (Accessed 05.10.2016)

GNS (2016). NGA GEOnet Names Server. http://geonames.nga.mil/gns/html/ (Accessed 05.10.2016)

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. http://dx.doi.org/10.1007/s10708-007-9111-y.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*(May), 110–120. http://dx.doi.org/10.1016/j.spasta.2012.03.002.

Graham, M., & De Sabbata, S. (2015). Mapping information wealth and poverty: The geography of gazetteers. *Environment and Planning A*, *47*(6), 1254–1264. http://dx.doi.org/10.1177/0308518X15594899.

Graham, M., De Sabbata, S., & Zook, M. A. (2015). Towards a study of information geographies: (Im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, *2*(1), 88–105. http://dx.doi.org/10.1002/geo2.8.

Grossner, K., Janowicz, K., & Keßler, C. (2016). Place, period, and setting for linked data gazetteers. In J. R. Mostern, & H. Southall (Eds.), *Placing names: Enriching and integrating gazetteers*. Bloomington, IN: Indiana University Press.

Guo, Q., Liu, Y., & Wieczorek, J. (2008). Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, *22*(10), 1067–1090. http://dx.doi.org/10.1080/13658810701851420.

Guptill, S. C., & Morrison, J. L. (1995). *Elements of spatial data quality*. Elsevier Science Limited.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. http://dx.doi.org/10.1068/b35097.

Hastings, J. T. (2008). Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, *22*(10), 1109–1127. http://dx.doi.org/10.1080/13658810701851453.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's Heart: The dynamics of the location field in user profiles. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 237–246). New York, NY, USA: ACM CHI '11 10.1145/1978942.1978976

Hengl, T., & Reuter, H. I. (Eds.). (2008). *Geomorphometry: Concepts, software, applications. Developments in Soil Science*, *vol. 33*. Elsevier 772 pp.

Hess, B., Magagna, F., & Sutanto, J. (2014). Toward location-aware web: Extraction method, applications and evaluation. *Personal and Ubiquitous Computing*, *18*(5), 1047–1060. http://dx.doi.org/10.1007/s00779-013-0718-3.

Hill, L. L. (2000). Core elements of digital gazetteers: Placenames, categories, and footprints. *Research and advanced technology for digital libraries* (pp. 280–290). Springerhttp://link.springer.com/chapter/10.1007/3-540-45268-0_26

Hill, L. L. (2006). *Georeferencing: The geographic associations of information*. The MIT Press.

Janowicz, K., & Keßler, C. (2008). The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, *22*(10), 1129–1157. http://dx.doi.org/10.1080/13658810701851461.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Kessler, C., Maué, P., Heuer, J. T., & Bartoschek, T. (2009). Bottom-up gazetteers: Learning from the implicit semantics of geotags. *GeoSpatial semantics* (pp. 83–102). Springerhttp://link.springer.com/chapter/10.1007/978-3-642-10436-7_6

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Vancouver, Canada: Sage.

Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. *Workshop on geographic information retrieval, Sheffield, UK*.

Leidner, J. L., & Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, *3*(2), 5–11.

Leveling, J. (2015). Tagging of temporal expressions and geological features in scientific articles. *Proceedings of the 9th Workshop on Geographic Information Retrieval* (pp. 6: 1–6:10). New York, NY, USA: ACM GIR '15 10.1145/2837689.2837701

Moncla, L., Gaio, M., & Mustière, S. (2014). Automatic itinerary reconstruction from texts. *Automatic itinerary reconstruction from texts. 8728*. (pp. 253–267). Vienna, Austria: Springer International Publishinghttp://link.springer.com/10.1007/978-3-319-11593-1

Mostern, R., Southall, H., & Berman, M. L. (Eds.). (2016). *Placing names: Enriching and integrating gazetteers*. Indiana University Press.

de Oliveira, M. G., Campelo, C. E. C., de Souza Baptista, C., & Bertolotto, M. (2016). Gazetteer enrichment for addressing urban areas: A case study. *Journal of Location Based Services*, *10*(2), 142–159. http://dx.doi.org/10.1080/17489725.2016.1196755.

OS 1:50k gazetteer (2016). Ordnance survey 1:50 000 scale gazetteer. Retrieved from https://www.ordnancesurvey.co.uk/business-and-government/products/50k-gazetteer.html (Accessed on 05.10.2016)

Overell, S., & Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, *22*(3), 265–287. http://dx.doi.org/10.1080/13658810701626236.

Popescu, A., Grefenstette, G., & Moëllic, P. A. (2008). Gazetiki: Automatic creation of a geographical gazetteer. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 85–93). New York, NY, USA: ACM JCDL '08 10.1145/1378889.1378906

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Gaihua, F., et al. (2007). The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science*, *21*(7), 717–745. http://dx.doi.org/10.1080/13658810601169840.

Schmachtenberg, Max, Christian Bizer, and Heiko Paulheim. 2014. "Adoption of the linked data best practices in different topical domains." In *The Semantic Web – ISWC 2014*, 245–60. Lecture Notes in Computer Science 8796. Springer International Publishing. http://link.springer.com/chapter/10.1007/978-3-319-11964-9_16.

Sehgal, V., Getoor, L., & Viechnicki, P. D. (2006). Entity resolution in geospatial data integration. *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems* (pp. 83–90). New York, NY, USA: ACM GIS '06 10.1145/1183471.1183486

Smart, P. D., Jones, C. B., & Twaroch, F. A. (2010). Multi-source toponym data integration and mediation for a meta-gazetteer service. *Geographic Information Science* (pp. 234–248). Springerhttp://link.springer.com/chapter/10.1007/978-3-642-15300-6_17

Smith, B., & Mark, D. M. (2001). Geographical categories: An ontological investigation. *International Journal of Geographical Information Science*, *15*(7), 591–612. http://dx.doi.org/10.1080/13658810110061199.

Smith, D. A., & Mann, G. S. (2003). Bootstrapping toponym classifiers. *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1* (pp. 45–49). Stroudsburg, PA, USA: Association for Computational Linguistics HLT-NAACL-GEOREF '03 10.3115/1119394.1119401

Southall, H., Mostern, R., & Berman, M. L. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, *5*(2), 127–145. http://dx.doi.org/10.3366/ijhac.2011.0028.

SwissNames 3D (2016). Retrieved from http://www.mont-terri.ch/internet/swisstopo/en/home/products/landscape/swissNAMES3D.html (Accessed 05.10.2016)

TGN (2015). Getty Thesaurus of geographic names: About the TGN. Retrieved from http://www.getty.edu/research/tools/vocabularies/tgn/about.html (Accessed 05.10.2016)

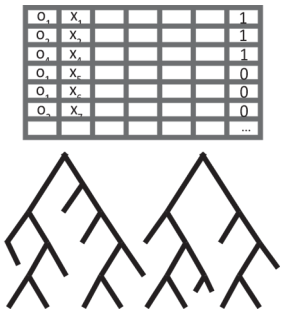
TGN (2016). Getty thesaurus of geographic names online. Retrieved from http://www.getty.edu/research/tools/vocabularies/tgn/ (Accessed 05.10.2016)

Van Oort, P. (2005). *Spatial data quality: From description to application*. Delft: Netherlands Geodetic Commission.

Yin, J., Karimi, S., & Lingad, J. (2014). Pinpointing locational focus in microblogs. *Proceedings of the 2014 Australasian Document Computing Symposium* (pp. 66:66–66:72). New York, NY, USA: ACM ADCS '14 10.1145/2682862.2682868

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, *9*(December). http://dx.doi.org/10.5311/JOSIS.2014.9.170.

Zhu, R., Hu, Y., Janowicz, K., & McKenzie, G. (2016). Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS*.



Paper II

Machine learning for cross-gazetteer matching of natural features

Summary

This paper presents a detailed methodology to perform cross-gazetteer matching using machine learning, implements this methodology using random forests, and applies it to an annotated dataset of natural feature records which we publicly released. Rule-based methods for gazetteer matching are also implemented and compared to machine-learning-based methods, and a thorough evaluation is carried out which highlights ways to construct a realistic testing pipeline and to avoid potential pitfalls which may optimistically bias generalization performance.

Contribution of the PhD candidate

I conceived of the idea for the paper, wrote all the analysis and plotting code in Python, drafted the manuscript, and prepared all the figures. Michele Volpi drafted part of the manuscript (random forests description in 'Machine learning based matching') and consulted on methods and results. Data annotation was done by Michele Volpi, Ekaterina Egorova, Julia Villette, and me (data released in a previous conference publication).

Citation

Acheson, E., Volpi, M., and Purves, R. S. (2019). Machine learning for cross-gazetteer matching of natural features. *International Journal of Geographical Information Science*, 1–27. DOI:10.1080/13658816.2019.1599123

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2019.1599123

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Machine learning for cross-gazetteer matching of natural features

Elise Acheson[a], Michele Volpi[b] and Ross S. Purves[a]

[a]Department of Geography, University of Zurich, Zurich, Switzerland; [b]Swiss Data Science Center, ETH Zurich and EPFL Lausanne, Switzerland

ABSTRACT

Defining and identifying duplicate records in a dataset is a challenging task which grows more complex when the modeled entities themselves are hard to delineate. In the geospatial domain, it may not be clear where a mountain, stream, or valley ends and begins, a problem carried over when such entities are catalogued in gazetteers. In this paper, we take two gazetteers, GeoNames and SwissNames3D, and perform matching – identifying records in each that are about the same entity – across a sample of natural feature records. We first perform rule-based matching, establishing competitive results, then apply machine learning using Random Forests, a method well-suited to the matching task. We report on the performance of a wider array of matching features than has been previously studied, including domain-specific ones such as feature type, land cover class, and elevation. Our results show an increase in performance using machine learning over rules, with a notable performance gain from considering feature types, but negligible gains from other specialized matching features. We argue that future work in this area should strive to be more reproducible and report results on a realistic testing pipeline including candidate selection, feature extraction, and classification.

ARTICLE HISTORY
Received 30 April 2018
Accepted 20 March 2019

KEYWORDS
Gazetteer matching; record linking; random forest; natural features; feature types

1. Introduction

Defining and identifying duplicate records in datasets is an important and persistent problem in an age of increasing quantities of heterogeneous digital data, produced by diverse methods ranging from crowdsourcing to expert curation. Geographical datasets present unique challenges stemming from the vague nature of many geographical entities, the high degree of referent ambiguity in geographical names, and the varied categorization systems for entity types in this domain. These conceptual challenges manifest themselves in how geographical entities are catalogued in gazetteers, resources storing minimally geographical names, types, and geometries for a defined region of interest (Hill 2006).

Indeed, different gazetteers (or more broadly, geospatial databases) can have, for a particular region of interest, the same entity listed under different names (e.g. Lake

CONTACT Elise Acheson elise.acheson@geo.uzh.ch

Geneva vs. Le Léman), represented with different geometries (e.g. two different point centroids, or a point and a polygon), and assigned to different feature types, from different feature type hierarchies (e.g. mountain vs. Haupthuegel, German for 'main hill'). These representational issues are exacerbated when dealing with natural features[1] such as mountains and valleys, which may have vague or varying extents, and name matching is made more difficult when dealing with multilingual data. Furthermore, the number and type of entities listed in different gazetteers can vary greatly, leading to resources with orders of magnitude more records than others for a given area – that is, with higher spatial coverage (Ahlers 2013, Acheson *et al*. 2017a).

Duplicate detection is a well-studied problem that has cut across disciplinary boundaries, and is thus, somewhat ironically, associated with a variety of names, including deduplication, entity resolution, and record linking (as also noted by Elmagarmid *et al*. (2007)). In GIScience, when two or more geospatial datasets are being aligned, the process is widely referred to as matching (e.g. Walter and Fritsch 1999, Olteanu *et al*. 2006, McKenzie *et al*. 2014, Morana *et al*. 2014). In the specific context of placename resources (gazetteers), we thus refer to record linking as *gazetteer matching*. Gazetteer matching aims to identify records referring to the same real-world geographical entity, to then potentially merge or integrate these co-referential records while still presenting a coherent and consistent picture of the world.

Research on gazetteer matching has so far been relatively sparse, and standardized methodology, tools, and reference datasets have yet to be firmly established. Nonetheless, published methods often share approaches considering place names, geometries, and optionally feature types, to triage records and identify duplicates, either using hand-crafted rules (Fu *et al*. 2005, Hastings 2008, Smart *et al*. 2010, McKenzie *et al*. 2014) or machine learning (Sehgal *et al*. 2006, Zheng *et al*. 2010, Martins 2011, Gonçalves 2012, Gelernter *et al*. 2013). However, details about the datasets used are often hard to come by, as are the datasets themselves, and comparisons between rule-based methods and machine learning approaches are largely absent. Furthermore, the focus has been primarily on coarse granularity feature types such as cities (Martins 2011, Gonçalves 2012), and on finer granularity urban feature types such as points of interest (Zheng *et al*. 2010, Gelernter *et al*. 2013, McKenzie *et al*. 2014).

Natural features are a largely neglected subset of geographical records in matching tasks, despite being considered prototypical 'geographic features' by many (Smith and Mark 2003) and clearly presenting the aforementioned challenges of vagueness and diverse type classifications.

In this paper, we align a subset of natural feature records from a global gazetteer, GeoNames, to records from an authoritative gazetteer for our study area, Switzerland. Through this process, we make the following contributions:

- We use open datasets and, for a subset of gazetteer records, a publicly available annotated gold standard (Acheson *et al*. 2017b), thus enabling future work to be directly comparable.
- We implement machine learning methods for the matching task and compare these to rule-based methods.

- For a given machine learning model, we test a wider array of matching features than has been previously studied, considering domain-specific ones such as feature type, land cover class, and elevation.
- We provide a full pipeline evaluation and highlight the importance of creating a realistic testing pipeline to obtain representative performance. Indeed, we show how machine learning performance can be artificially affected by the choice of positive and negative record pairs.

In what follows, we motivate these contributions through a review of relevant work, particularly gazetteer matching and its methodological components in a rule-based and machine learning context (section 2). We then describe the two gazetteers used and our annotation process in section 3, followed by details of our rule-based matching methods and machine learning based methods in section 4. In section 5, we present the results of our approaches to automatically find matches between the two gazetteers. Our subsequent discussion centers around the many facets of matching that impact performance, and based on our detailed analysis we conclude with recommendations for future work.

2. Related work

Gazetteer matching is a special case of the widely studied problem of record linkage and part of the broader challenge of entity resolution (Elmagarmid *et al*. 2007, Costa 2011, Christen 2012). In gazetteer matching, the records to be linked represent geographical entities, rather than people, biomedical records, or web pages. The vast amounts of literature on record linkage, produced by various research communities, testify to the ubiquity of the problem, and to the added value of good solutions. As the need for gazetteer matching arises from heterogeneously catalogued data, we first briefly discuss the properties and uses of gazetteers, our data sources. We then focus on literature specifically concerned with gazetteer matching, and situate these works where appropriate within the broader context of entity resolution.

2.1. Gazetteers

Gazetteers, geographical datasets cataloguing named places alongside their feature types and geometries (Hill 2006, Berman *et al*. 2016), are important resources for linking unstructured textual content to geographical space. By providing explicit spatial representations (geometries such as points, lines, and polygons) for geographical references used in natural language (placenames, also known as toponyms), they can serve as 'glue' to explicitly spatialize textual data of various types, thus opening this data up for spatial analyses. Tasks that make use of gazetteers include detecting placenames in text, disambiguating and grounding placenames, and retrieving information about named places such as feature type, population, and containment relationships with other places or regions (Purves *et al*. 2007, Lieberman *et al*. 2010, Cooper and Gregory 2011, Adams *et al*. 2015).

Gazetteers are necessarily simplified versions of a more complex geographical reality, where continuous space is neatly carved up into objects and packaged into rows with attributes following a given schema. Data heterogeneity is all but unavoidable and can exist on several levels, including:

- **names**: places often have multiple names (whether spelling variants or across languages), may include fossilized type information within the name (e.g. Lake Placid), and the same name can be used for multiple entities, often in close proximity (e.g. Lake Placid the town and Lake Placid the lake) (Brunner and Purves 2008, Hastings 2008).
- **geometries**: the geometries representing named places can be different (e.g. two different points), of a different type (e.g. a point and a line), and of varying geometric complexity.
- **feature types**: each gazetteer can have its own feature type hierarchy used to classify real-world entities, and within a hierarchy, different types could be assigned to gazetteer records representing the same entity.

With the vast quantities of geotagged data available today online, particularly produced by non-experts and pushed to various social media or photo-sharing platforms, gazetteer production processes have evolved to include not just top-down, curated resources, but 'bottom-up' gazetteers from crowdsourced data (Popescu *et al*. 2008, Keßler *et al*. 2009, Gao *et al*. 2017). This means more resources exhibiting potentially high levels of heterogeneity and requiring both internal deduplication and robust record linking across datasets.

2.1.1. Feature type hierarchies

Categories of geographical entities are useful for communication and reasoning in everyday situations, such as when describing a hike 'through a **valley**' or 'up a **mountain**'. In the context of gazetteers, these categories are known as feature types, and feature type categorization systems are referred to varyingly as feature type hierarchies (our preferred term), schemes, thesauri, or ontologies, depending on their characteristics (e.g. see Janowicz and Keßler (2008)). Feature type hierarchies are one important way in which gazetteers vary, since these hierarchies tend to be idiosyncratic, with types appropriate for a gazetteer's area of coverage, language(s), and purpose (Hastings 2008, Janowicz and Keßler 2008).

Works dealing with aligning feature type hierarchies in a gazetteer matching context face a 'chicken or egg' problem: the feature types of records may be used as evidence that two records are about the same entity, and linked records may in turn be used to calculate feature type similarity. Taking a pragmatic approach, Brauner *et al*. (2007) calculate similarity measures between feature types in two gazetteers based on records deemed to be about the same entity, but offer few details on this record linking process, which relies heavily on geographic location alone. Hastings (2008) manually conflates different feature type hierarchies into one in order to then perform gazetteer matching. Smart *et al*. (2010) follow a similar path, developing a custom feature type ontology based on their gazetteer data sources and suitable for their tasks. In a machine learning based matching context, Sehgal *et al*. (2006) use annotated record pairs to calculate a static type similarity metric between feature type pairs, then used in classification. In a work concerned with type alignment rather than gazetteer matching, Zhu *et al*. (2016) calculate spatial statistics for records of a subset of feature types to derive 'spatial signatures' for each type, which they argue may complement existing type alignment methods.

2.2. Gazetteer matching

Entity resolution in general consists of a sequence of steps (Figure 1), usually comprising a *data preparation* step (establishing which fields are to be compared and cleaning and normalizing records), a *record linking* step (matching corresponding records), and potentially a *record fusion* step (merging/augmenting records deemed to be about the same entity) (Elmagarmid *et al*. 2007, Costa 2011). The second step, record linking, is the focus of this paper and, in the context of our work, we refer to it as gazetteer matching. Thus, gazetteer matching consists of linking pairs of gazetteer records which are thought to refer to the same real-world entity. A special case of the problem is deduplication, where duplicate records in a single gazetteer are identified and merged (Christen 2012). Cross-gazetteer matching, however, presents additional challenges compared to deduplication, including dealing with multiple feature type hierarchies (Janowicz and Keßler 2008), with structural heterogeneity such as different schemas (Elmagarmid *et al*. 2007), and with varying spatial coverage and balance between gazetteers (Acheson *et al*. 2017a).

Broadly, matching methods can be divided into rule-based (or distance-based) approaches and machine learning (or probabilistic) approaches (Elmagarmid *et al*. 2007). Rule-based approaches rely on either a series of binary rules for triaging records until only matches remain, or on setting weights manually to a set of distance (or similarity) measures to identify the most similar record(s) for each candidate. Rule-based methods can be considered a special case of distance-based methods, where distances are boolean (Elmagarmid *et al*. 2007). Machine learning approaches treat the matching problem as a supervised binary classification problem, outputting a 'match' or 'no match' label for each candidate pair of records. To train such a classifier, a set of matching record pairs (positive training examples) and non-matching pairs (negative training examples) are required. The key difference between rule-based and machine learning methods is thus whether decision boundaries between matches and non-matches are defined heuristically or in a data-driven fashion, with the latter explicitly requiring training data. Generally, machine learning methods are considered to offer more flexible and performant solutions, with the caveats of being more complex to implement and more time-consuming due to the need for annotated training data. Correspondingly, the main advantages of rule-based methods over machine learning methods are their simplicity and the reduced need for annotated data, since many rule-based methods only use data in evaluation, but not training (Elmagarmid *et al*. 2007). For gazetteer matching, this is an appealing advantage due to the paucity of annotated

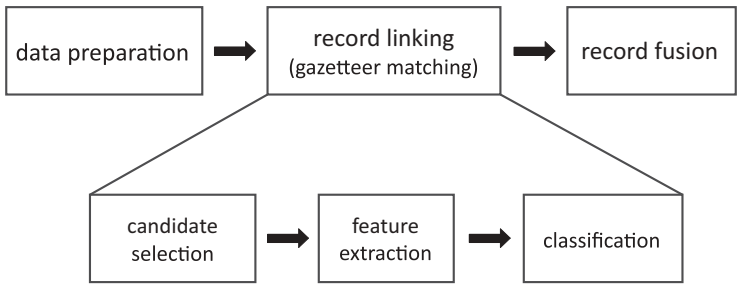


Figure 1. Entity resolution steps and sub-steps.

datasets, despite some recent efforts, including a benchmark building tool (Morana *et al*. 2014) with restrictive target datasets, and a public, but not cross-gazetteer, dataset (Gonçalves 2012).

2.2.1. Gazetteer matching steps

Gazetteer matching can itself be divided into steps (Figure 1). Whether one takes a rule-based or machine learning approach, the first step to finding matching records is to find match candidates, known as *candidate selection* (Zheng *et al*. 2010). Indeed, for any reasonably large dataset, considering every pair of records as a match candidate is infeasible and unnecessary: geographic records with locations that are far apart are improbable matches. This candidate selection step is also known as 'filtering' (Martins 2011), 'blocking' (Elmagarmid *et al*. 2007, Morana *et al*. 2014), or 'indexing' (Christen 2012). The general idea is, for every source record, to select from all target records only a subset of likely match candidates. For example, Zheng *et al*. (2010) select candidates using 'simple heuristics' including ignoring records with locations too far apart or with completely dissimilar names. In practice, candidate selection can consist of an initial coarse triage of target records, followed by a more careful selection of candidates via secondary filtering criteria or a ranked list. In one such case, Morana *et al*. (2014) use a feature-type specific point-radius method as an initial spatial filter alongside a type filter, then retain any candidate that shares a token with the name to match (e.g. New Amsterdam vs. New York), a pragmatic requirement for their monolingual points of interest (POI) data in a language which does not use compound nouns.

Once matching candidates are identified for a particular record, distance (or similarity) metrics can be calculated between each remaining pair of records. In the context of machine learning, these metrics are called 'matching features', and thus this step is known as *feature extraction* (Zheng *et al*. 2010). Pairwise metrics are typically calculated on place names, locations (geometries), and feature types (Sehgal *et al*. 2006, Zheng *et al*. 2010, Martins 2011). In particular, many algorithms exist which estimate name similarity, including character-based, token-based, and phonetic algorithms (Elmagarmid *et al*. 2007). Sehgal *et al*. (2006) find that the character-based Levenshtein distance (Levenshtein 1966) (also known as edit distance) is the best similarity metric for location names, but Zheng *et al*. (2010) argue that custom token-based metrics are more suitable for their POI and address data. Smart *et al*. (2010) use a combination of the Levenshtein distance, text normalisation, and the SoundEx phonetic algorithm. Martins (2011), extended in Gonçalves (2012), explore a wide variety of name similarity metrics for gazetteer record deduplication, including the Levenshtein, Jaro-Winkler, Monge-Elkan, Double Metaphone, and Jaccard algorithms. As for matching features on feature types, options vary based on whether types are from a single hierarchy and whether records can have more than one type assigned. With one hierarchy and multiple types per record, matching features can be derived based on the hierarchical distance between the types and on the overlap between assigned types (Zheng *et al*. 2010, Martins 2011). With multiple typing systems, options include manually aligning relevant types (Morana *et al*. 2014) or estimating type alignment based on annotated records (Sehgal *et al*. 2006).

After selecting candidate pairs and calculating matching features for each pair, a final *classification* step outputs a decision for each pair as to whether they form a match or not. Using rules, such a decision can be made by manually setting a threshold on an overall similarity score or on individual scores, whereas using machine learning, decisions are

probabilistic and depend on the algorithm and training data used to build a predictive model. Machine learning algorithms used for the gazetteer matching task have included Logistic Regression, Voted Perceptron (Neural Network), Support Vector Machines (SVMs), Decision Trees, and Random Forests (Sehgal *et al*. 2006, Zheng *et al*. 2010, Martins 2011, Gonçalves 2012). In particular, random forests have been found to outperform both decision trees and SVMs in a gazetteer deduplication context (Gonçalves 2012). Deep learning has not yet been used in gazetteer matching, but Santos *et al*. (2018) use deep learning in the form of a recurrent neural network to classify whether pairs of placenames are in fact alternate names for the same geographic entity, using the large GeoNames gazetteer as training data. However, deep learning requires very large annotated subsets for training, a property which is not likely to be satisfied in most heterogeneous gazetteer matching tasks.

2.2.2. Training and testing in machine learning based matching

A particular challenge in machine learning based gazetteer matching is the selection of training and test data. How does the size and composition of the training data impact classification performance? In particular, what ratio of matches (positive training examples) to non-matches (negative training examples) should the training data comprise and how should non-matches be selected? How should test data be selected and processed and will performance on this test data generalize to a wider dataset?

Sehgal *et al*. (2006) investigate in some detail how to choose non-matches for training in a gazetteer matching context, settling on a combination of random non-matching pairs and 'hard negatives', where hard negatives are non-matching records that have either highly similar names or highly similar locations. They choose the top k hard negatives per record, where k is optimized experimentally based on classification performance. Their highest performing algorithm (Logistic Regression) and training set composition consists of 30 negative training examples for every positive example. Martins (2011) and Gonçalves (2012) follow a similar procedure but opt for a 1:1 ratio of non-matches to matches, additionally specifying that half of their non-matches are selected randomly, while the other half comes from a ranked list of hard negatives for the whole collection. In Zheng *et al*. (2010), few details are given as to how non-matches were selected, but their overall dataset consists of a 1:1 ratio of matches and non-matches (800 each).

In addition, cross-validation is often the end point of a machine learning classification pipeline, with no independent, unseen test set. Since it is well known that unbalanced datasets and poorly chosen evaluation data can lead to wildly overoptimistic (or indeed pessimistic) evaluations in a wide range of contexts (Murphy 1996), we argue that there is a need for increased clarity in how training and testing data are chosen and processed in gazetteer matching.

3. Data

3.1. Application context

Our specific motivation for performing gazetteer matching is to aid in the georeferencing of Swiss alpine journal texts. These texts consist largely of descriptions of ski tours and hikes, and thus contain textual references to natural features such as mountains, valleys, and glaciers. Most of the natural features mentioned in these texts are located in

Switzerland, but some are in other mountainous parts of the world. Thus we use both SwissNames3D, an official placename resource for Switzerland, and GeoNames, an unofficial global resource. We wish to reconcile overlapping records for a cleaner georeferencing process, by linking GeoNames records to their SwissNames3D equivalent(s) in order to potentially increase recall while maintaining precision.

3.2. Gazetteers

SwissNames3D is an authoritative gazetteer of placenames in Switzerland, which is freely downloadable online.[2] A new edition is published annually, with the full country-wide update cycle taking 6 years. We downloaded the full dataset in February 2017. It contains over 300k records, each with a unique identifier, organized according to a Switzerland-specific feature type hierarchy, with lines and polygons available for a large subset of records depending on their types (e.g. streams are available as lines and valleys are available as polygons).

GeoNames is a widely used global gazetteer composed from a variety of sources including open geographical datasets and user-contributed data. Because of its global coverage and easy availability, GeoNames is very widely used, though it has also been recognised that its bottom-up production makes understanding data quality challenging (Ahlers 2013, Acheson *et al*. 2017a). We downloaded the freely available, daily updated GeoNames data for Switzerland on 20 July 2017.[3] It contains around 67k records for the country, organized according to a two-tiered, global feature type hierarchy, and with points available for all records in this free version.

3.3. Annotation

We manually prepared an annotated gold standard for a portion of records in GeoNames. Since SwissNames3D is an official resource and contains a much larger number of records than GeoNames for Switzerland, we assumed it to be more accurate and complete, thus better suited as our target resource for matching. We thus started with our source dataset GeoNames, and retained all feature types that we identified as representing natural (as opposed to human-made) geographic features and having at least 100 records in Switzerland. From the 8 types that met these criteria, we randomly selected 50 records of each type for annotation (see Table 1).

Table 1. Selected natural feature types from GeoNames, along with a representative type in SwissNames3D, and their counts.

GeoNames			SwissNames3D	
type	count	annotated	type	count
lake (LK)	1132	50	See	1263
glacier (GLCR)	806	50	Gletscher	854
stream (STM)	172	50	Fliessgewaesser	6603
peak (PK)	6557	50	Gipfel	2225
pass (PASS)	1785	50	Pass	2290
hill (HLL)	665	50	Huegel	1840
mountain (MT)	352	50	Haupthuegel	938
valley (VAL)	113	50	Tal	2260

For these 400 GeoNames records, one annotator per record tried to comprehensively find matches in SwissNames3D, including one-to-many and 'no match' cases. Any harder cases were then discussed among the four annotators, all graduate students in Geographic Information Systems, until an agreement was reached (as described in Acheson *et al*. 2017b). These annotated data provide us with training and evaluation data for matches, but not non-matches, in the gazetteer matching experiments which we now describe.

4. Methods

We implemented and compared rule-based matching and machine learning based matching. In terms of the overall entity resolution pipeline laid out in Figure 1, in both cases we performed the *data preparation* step manually, then focused on the core *record linking/gazetteer matching* step and its sub-steps. *Record fusion* (merging/augmenting linked records) was not performed for this work, but as SwissNames3D is an authoritative resource, fusion could consist of adding information from GeoNames which is not present in SwissNames3D, such as some alternate names. Data preparation included identifying which fields should be compared, and projecting the GeoNames latitude and longitude coordinates (WGS84) to Swiss coordinates (LV03) and vice-versa, to facilitate distance calculations. Additional work was required to deal with a peculiarity of SwissNames3D, which uses multiple records, each with the same table ID, for a single geographic entity when this entity has an official name in more than one official language of Switzerland. To get around this issue, we created a truly unique ID for each record, then ran our entire pipeline treating every record as unique, before reconstructing the original IDs for evaluation in order to not underestimate recall.

In the remainder of this section, we give an overview of our matching features, then present our rule-based matching methods. We then focus on our machine learning matching pipeline and how we implemented each step, and finally present our evaluation methods.

4.1. Matching features overview

Previous work has consistently used record names and geometries in gazetteer matching, firmly establishing their utility. Consequently, our rules and machine learning feature combinations always consider names, minimally as the Levenshtein distance between pairs of record names, and geometries, as the point-to-point distance between records (Vincenty 1975); the only exception is our rule-based random baseline which only considers names. As for feature types, their use and treatment is less consistent in the literature, with many derived matching features requiring a single feature type hierarchy. In this work we use feature types in two main ways: as an initial filter to limit the number of target records we consider during candidate selection, and as categorical features for machine learning during feature extraction. For candidate selection, we first established soft type alignments (Table 2) between the types of interest in our two different feature type hierarchies (a Switzerland-specific hierarchy with types in German, and a global hierarchy with types in English), using feature type metadata rather than our annotated data. However during feature extraction for machine learning,

Table 2. Soft type alignments used and type-specific distance thresholds.

GeoNames type	SwissNames3D types	Threshold (km)
mountain (MT)	Hauptgipfel, Gipfel, Huegel, Haupthuegel, Alpiner Gipfel, Grat, Huegelzug, Felskopf	5
hill (HLL)	Hauptgipfel, Gipfel, Huegel, Haupthuegel, Alpiner Gipfel, Grat, Huegelzug, Felskopf	5
peak (PK)	Hauptgipfel, Gipfel, Huegel, Haupthuegel, Alpiner Gipfel, Grat, Huegelzug, Felskopf	5
glacier (GLCR)	Gletscher, Alpiner Gipfel	5
pass (PASS)	Pass, Graben	5
lake (LK)	See, Seeteil	15
stream (STM)	Fliessgewaesser	15
valley (VAL)	Tal, Haupttal, Graben	15

we do not use these alignments, and instead encode the categorical feature type information as a set of binary features labelling the presence or absence of a given type, a common strategy known as 'one-hot encoding'. We also indirectly use feature types in one of our rule-based procedures to define type-specific geographical distance thresholds. In addition to names, geometries, and types, we also make use of elevation and land cover information as a way to represent properties of the natural environment, which may be useful in matching our natural feature records. A detailed list of all our matching features is given in 4.3.3.

4.2. Rule-based matching

We implemented rule-based methods to obtain competitive results while also testing the utility of matching features in a deterministic way. We tested the following rule-based matching procedures, in order of increasing complexity:

- **random-baseline**: find all exact name matches on the primary name for a given source record, then randomly choose one exact match as the match (no exact name matches means no match).
- **name-threshold**: find all exact name matches on the primary or any alternate name for a given source record, then from these, retain all target records within a fixed distance threshold (e.g. 5km) of the source record; here, the set of results can have 0, 1, or multiple matches per source record.
- **name-custom-threshold**: proceed as in *name-threshold* above, but this time use custom thresholds (see Table 2) specific to the feature type of the source record (c.f. Morana *et al*. 2014).
- **multi-threshold**: proceed as in *name-custom-threshold* above, but discard any target records above a threshold on land cover distance (as described in 4.3.3) or elevation.
- **linear-combination**: find all exact name matches on the primary or any alternate name as before, and additionally retain any target records of an aligned feature type (see Table 2) for each source record, then calculate edit distance (Levenshtein) and geographical distance for each pair. Combine these two distances in a weighted sum for a final score and keep any pairs with a score above an empirically derived threshold as a match (c.f. Smart *et al*. 2010).

Our rule-based procedures above combine matching features either by sequentially applying thresholds (*name-threshold, name-custom-threshold, multi-threshold*) or by including them in a linear combination (*linear-combination*). Since an important real-world advantage of rules over machine learning is simplicity, in terms of both interpretability and development time, we spent some time optimizing thresholds (for geographical distance and in *linear-combination*, for the overall score), but did so manually by considering sensible values for distance thresholds and by testing values at fixed intervals for the overall score. Furthermore, the more matching features are used, the more thresholds or weights there are to optimize overall, which incentivizes a sparser use of matching features. As a compromise between using a minimum number of matching features and having to use machine-learning-like techniques to combine a large number of features, we include a procedure (*multi-threshold*) which combines at least one matching feature from each of the five categories (names, geometries, types, elevation, and land cover).

4.3. Machine learning based matching

We frame machine learning based matching as a binary classification problem. Our aim is to build a model to infer whether a pair of records is a match or not, that is, whether they refer to the same real-world entity. To this end, we use a Random Forest classifier (Breiman 2001). We selected this classifier for several reasons. First, its simplicity: Random Forest is a non-linear, non-parametric classifier, which is intrinsically regularized by ensembling and not prone to overfitting. The key parameter choice in a random forest is the number of trees used in the ensemble: the larger the number, the less prone to overfitting the model is. We settled on 200 trees after finding that performance plateaued around this number. Second, as opposed to most probabilistic and distance-based classifiers (e.g. non-linear Support Vector Machines, Naive Bayes, etc.), input features do not have to be pre-processed to follow some particular distribution, nor do they have to be normalized. Random forests can also naturally handle categorical and continuous features jointly, a key advantage in our case to work with categorical features, such as feature types and land cover classes, alongside continuous features, such as geographical distance and elevation. As mentioned, categorical features are 'one-hot encoded', that is, M categories are encoded as M binary features labelling the presence or absence of a given category (e.g. category 4 out of 5 is thus encoded as [0,0,0,1,0], category 2 as [0,1,0,0,0] and so on). Finally, random forests have been successful in many applications, being better or at least on par with most non deep-learning classification algorithms.

In our implementation, we ask the random forest to infer whether a previously unseen pair of records (test data) is a match or not. To this end, the feature vector representing the pair is passed through every tree and the ensemble outputs a distribution over the labels. The final solution is given by taking the maximum-a-posteriori over the predicted posterior.

4.3.1. Pipeline overview

We built a machine learning processing pipeline (Figure 2) with the aim to approximate a realistic, large-scale matching scenario. First of all, we assume the dataset is too large to calculate matching features for every possible record pair, and thus the

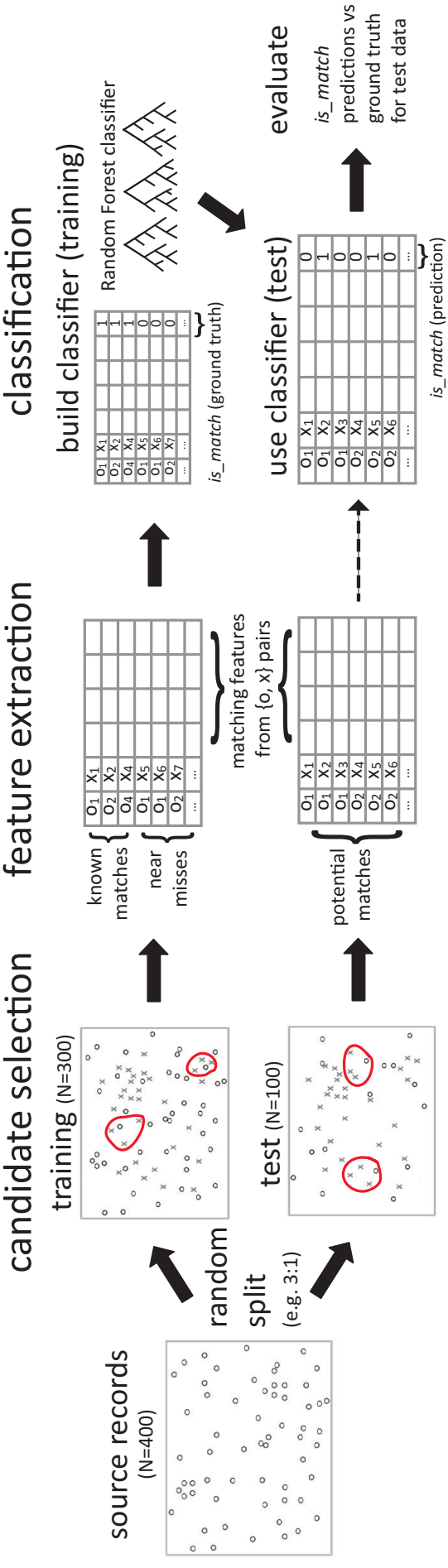


Figure 2. Machine learning pipeline overview.

first matching step must be to choose a subset of pairs, via candidate selection. Second, we assume that true positive pairs are not known in advance for the test set, as for truly unseen data, and thus positive matches in our test set must all be found through candidate selection. Indeed in our testing pipeline, we only classify pairs with target records retained at the candidate selection stage. In our training pipeline however, we ensure that the full set of annotated matching pairs for our source training records gets used, allowing our classifier to be optimised based on all of the information we have available. In order to process the training and test set in this slightly different manner, we split the source records at the beginning of a run, prior to candidate selection.

A single run of the pipeline takes a portion of source records through the training pipeline to train a classifier, then evaluates this classifier on the remaining, held-out, source records which go through the testing pipeline. As a compromise between maximizing the size of the training data and having a large enough test set, we opt to split the full set of source records ($N = 400$) randomly as 75% training and 25% test for our main runs. We perform 20 runs with this splitting procedure, each time testing our full set of feature combinations (section 4.3.4) on this particular split, to obtain not only mean values of precision, recall, and F1 per classifier, but also interquartile ranges (results in section 5.2). As we do not constrain the random split, the test sets vary in composition and difficulty, which is desirable to ensure robustness in a real-world scenario. We also create a fixed, feature-type-balanced test set, consisting of 10 randomly chosen source records from each of our 8 GeoNames types, for a total of 80. This fixed test set allows us to compare the performance of the rules and machine learning methods evaluated on the exact same data subset. We also use this fixed test set to plot a learning curve for different feature combinations, that is, to show how performance changes as we use more and more training data to train the random forest (results in section 5.3).

All the processing code was implemented in python, relying primarily on the *pandas*[4] library to work with data tables and compute matching features, and the *scikit-learn* library for machine learning (Pedregosa *et al*. 2011). We make our code and analysis files publicly available.[5]

4.3.2. Candidate selection

The first gazetteer matching step is candidate selection. For the training set, the aim of candidate selection is primarily to find (hard) negative record pairs to go alongside the known positive record pairs in order to train a successful classifier, and to make the subsequent machine learning independent from the absolute size of the data to be matched. For the test set however, candidate selection should aim to retain as many positive record pairs as possible (alongside some negative pairs), since the true positives would not be known in advance in a real matching scenario. Thus theoretically different candidate selection methods could be used for the training and test set, especially to try to increase recall on the test set, since any true positive pairs not retained during candidate selection will simply not make it to the classification step.

We selected candidate matches in the same way for both training and test pipelines. After a loose feature type filter (Table 2), we calculate the Levenshtein distance on names and the point-to-point distance on geometries, then combine these for an overall

score, similar to our *linear-combination* procedure. We then retain the top k candidates per source feature, where k was set experimentally by comparing performance for a range of values of k (results in section 5.2).

4.3.3. Feature extraction

The second matching step is to compute a range of matching features between all of the candidate source-target record pairs for use in the random forest. We chose a wide range of features to capture the similarity of the record names, locations, feature types, and geographical context via elevation and land cover. For land cover data, we use nationally-produced data for Switzerland containing a top level 6-class categorization of land cover.[6]

Our full list of features is as follows:

- **names**: for primary names, we calculate the Levenshtein distance, the normalized Levenshtein-Damerau distance, the Jaro similarity, and the Jaro-Winkler similarity; we additionally calculate the Levenshtein distance on any alternate names (present in GeoNames only) and on names with a comma, where we remove the comma and move the token following the comma to the beginning; finally we also take as a feature the minimum Levenshtein distance of those calculated.
- **geometries**: we calculate the point-to-point distance between gazetteer records (Vincenty 1975).
- **feature types**: we use one-hot encoding to encode feature type information, which removes any need to manually align our differing feature type hierarchies.
- **elevation**: we calculate the absolute difference between elevation values associated with the placenames in each gazetteer (essentially a measure of relief).
- **land cover**: we derive three land cover features from the land cover data. First, we find the land cover class of the nearest cell for both the source and target record, then one-hot encode this class. Second, we find the most frequent land cover class of the nearest 9 cells for each record and again one-hot encode this class. Finally, we calculate a feature we call 'land cover distance', where we take the counts of the 6 land cover classes in the 9 nearest cells for both source and target record (for example $[0, 4, 3, 1, 0, 1]$ and $[0, 4, 2, 0, 0, 3]$) then take the sum of the absolute value of the difference between these arrays (for example $[0, 0, 1, 1, 0, 2]$ for the absolute difference and $1 + 1 + 2 = 4$ for the sum, our final numeric feature).

4.3.4. Classification

We tested various combinations of the matching features we describe above, guided by our literature review. Based on reported strong performance of the Levenshtein distance for location name matching (Sehgal *et al*. 2006, McKenzie *et al*. 2014), we first formed a *basic* model using the minimum Levenshtein distance and the point-to-point distance. As a variant of the *basic* model, we formed a *str* model where we included all of our name matching features alongside point-to-point distance. We then added feature types to these two models, forming our *basic-type* and *str-type* models. We formed *str-elev-lc* to test whether we could compensate for a lack of feature type information by using elevation and land cover features. Finally, we tested 3 variants where most or all features are present, including feature types (*str-type-lcd, all-min, all*).

The feature combinations we focused on are thus as follows:

- **basic**: minimum Levenshtein distance and geographical distance.
- **str**: all name (string) features and geographical distance.
- **basic-type**: minimum Levenshtein distance, geographical distance, and encoded feature types.
- **str-type**: all name (string) features, geographical distance, and encoded feature types.
- **str-elev-lc**: all name (string) features, geographical distance, elevation, and all land cover features (no feature type information).
- **str-type-lcd**: all name features, geographical distance, encoded feature types, and land cover distance.
- **all-min**: minimalist version still using one feature per category: minimum Levenshtein distance, geographical distance, encoded feature types, elevation, and land cover distance.
- **all**: all features.

4.4. Evaluation

We evaluate the performance of both our rule-based and machine learning based matching using standard precision, recall, and F1 measures (Sehgal *et al*. 2006). Precision is defined as the number of positive matches correctly found divided by the total number of positive matches found, while recall is defined as the number of positive matches correctly found divided by the total number of positive matches that were to be found. Since precision can typically be optimized at the expense of recall and vice-versa, F1 is a measure combining precision and recall through their harmonic mean and thus summarizes the overall performance.

In our case, there is an additional complexity to be aware of with respect to recall. Calculating recall requires knowing how many positive matches there were in total involving the source records in the test set. Some of these positive matches will however not make it through the candidate selection stage, such as pairs with both dissimilar names and locations. Since in our test pipeline, we only keep candidate matches that were retained in candidate selection, we have two sets of matches we can use as the recall denominator: the full set of matches involving our test records (*overall recall*) or just those matches that made it to the classification stage (*classification recall*).

We consider *overall recall* to be the more meaningful recall of the two, since in a real-world scenario, the full set of correct matches for each source record is not known in advance, but instead the correct target records have to be found via candidate selection. The *classification recall* however serves as a useful evaluation of the classification stage specifically, without considering directly how well or poorly candidate selection performed. We can calculate an upper bound for overall recall right after candidate selection by looking at the percentage of positive matches that we retained out of the full set of known positives for the test records. We refer to this upper bound as *max recall* and present it alongside the other values described.

5. Results and interpretation

We describe here the results of our matching experiments, first presenting results from our rule-based matching procedures, then detailing our results using random forests. To describe the performance of our random forests, we first show how we fixed *k*, the number of target record candidates per source record used for candidate selection. We then present our main results: precision, recall, and F1 performance over 20 runs for all our combinations of matching features. Finally, we report results obtained on our fixed, feature-type balanced test set and plot a learning curve to show how performance changes as we increase the size of the training set.

5.1. Rule-based matching

Our rule-based matching results are shown in Table 3, in order of increasingly complex rules. Results are shown for a distance threshold of 5km for *name-threshold*, type-specific thresholds of either 5km or 15km for *name-custom-threshold* and *multi-threshold* (see Table 2), and additional thresholds of 400m of elevation difference and 8 units of land cover distance for *multi-threshold*. For *linear-combination*, results are shown using a single overall threshold and two weighting schemes for textual and geographical distance, one for lakes, streams, and valleys (LK, STR, VAL) which gives lesser weight to geographical distance, and one for all other types which gives equal weight to both.

The relatively high precision of *random-baseline*, where we chose a random exact name match as the match, shows that a significant proportion of the data can be dealt with using names alone. For two more complex sets of rules, *name-custom-threshold* and *linear-combination*, we were able to obtain good overall performance, with F1 values reaching around 0.85. Our *multi-threshold* approach, which employs additional thresholds on elevation and land cover, can clearly increase precision, but at the cost of much lower recall, and overall lower F1 values. Indeed, the best F1 performance we obtained with *multi-threshold* after trying a range of values for the additional thresholds was simply to not have these thresholds, which is equivalent to *name-custom-threshold*.

The higher performance of *name-custom-threshold* and *linear-combination* come at the cost of having to set several parameters in an ad hoc fashion, including multiple distance thresholds for *name-custom-threshold* and an overall score threshold with type-specific weightings for *linear-combination*. This threshold and the weightings could be optimized further, particularly to favour precision over recall or vice-versa, depending on the task requirements, by setting up a grid-search using one of the performance measures detailed above and evaluating the performance on a held-out or cross-validation sample. However, both trying to combine many matching features and trying

Table 3. Results for rule-based matching.

name of run	precision	recall	F1
random-baseline[a]	0.793	0.575	0.666
name-threshold	0.876	0.788	0.830
name-custom-threshold	0.843	0.861	0.852
multi-threshold	0.914	0.677	0.778
linear-combination	0.871	0.833	0.852

[a]random-baseline results were averaged over 10 runs.

to fully optimize rule-based matching leads us quickly down a data-driven path for which machine learning is better suited.

5.2. Machine learning based matching

We now present our machine learning based matching results, starting with how we fixed a value for k, the number of target records retained per source record during candidate selection. Figure 3 shows the overall performance (F1) of our machine learning pipeline for different feature combinations and values of k. Values of k below 10 appear to decrease F1, while a value of 30 appears to be optimal for most of the feature combinations tested and was used in the subsequent experiments.

In Figure 4, we present our main results for the full machine learning pipeline, showing the performance of different feature combinations over 20 runs with $k = 30$, in terms of F1, precision, (overall) recall, and classification recall. As mentioned, classification recall is calculated on just those positive record pairs retained during candidate selection, whereas overall recall is calculated using the full set of annotated positive pairs. For each run, all feature combinations were trained and tested on a particular random data split. This means each feature combination was evaluated against the same 20 sets of test records, presenting mixed feature type profiles and variations in difficulty.

Overall, median F1 values start at around 0.80 for our *basic* feature combination and increase up to 0.90 for the *str-type-lcd* combination. Visible on the F1 plot is a clear

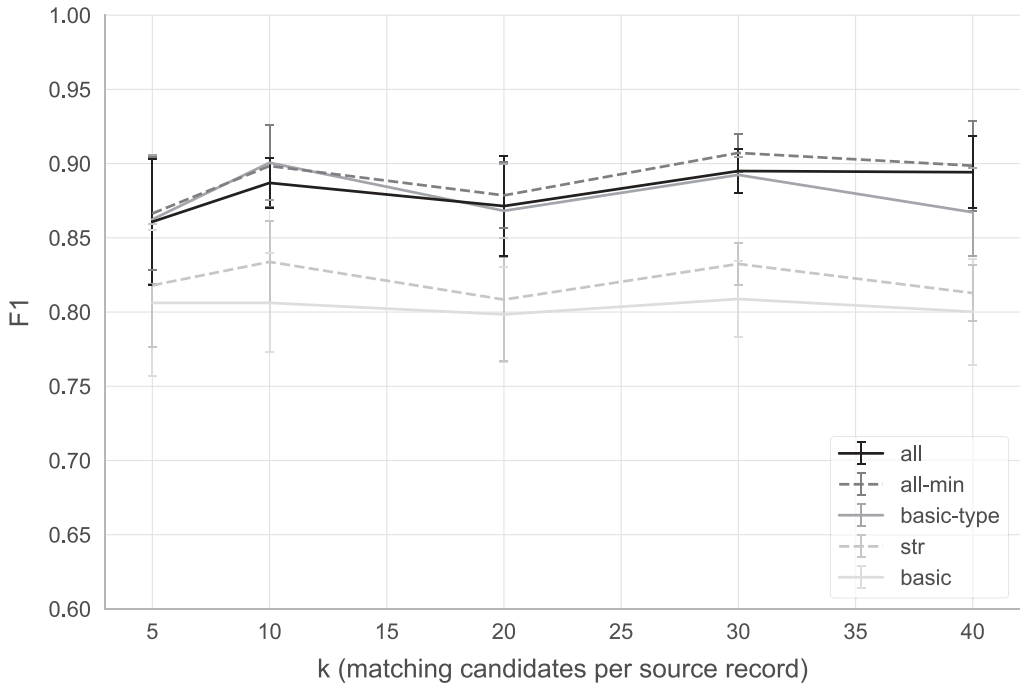


Figure 3. Mean and standard deviation of F1 vs k (number of target records retained per source record during candidate selection) over 5 runs per data point (tested over 5 values of k and 5 feature combinations). Feature combinations plotted: *basic*: minimum Levenshtein distance and geographical distance; *str*: all name (string) features and geographical distance; *basic-type*: minimum Levenshtein distance, geographical distance, and encoded feature types; *all-min*: minimum Levenshtein distance, geographical distance, encoded feature types, elevation, and land cover distance; *all*: all features.

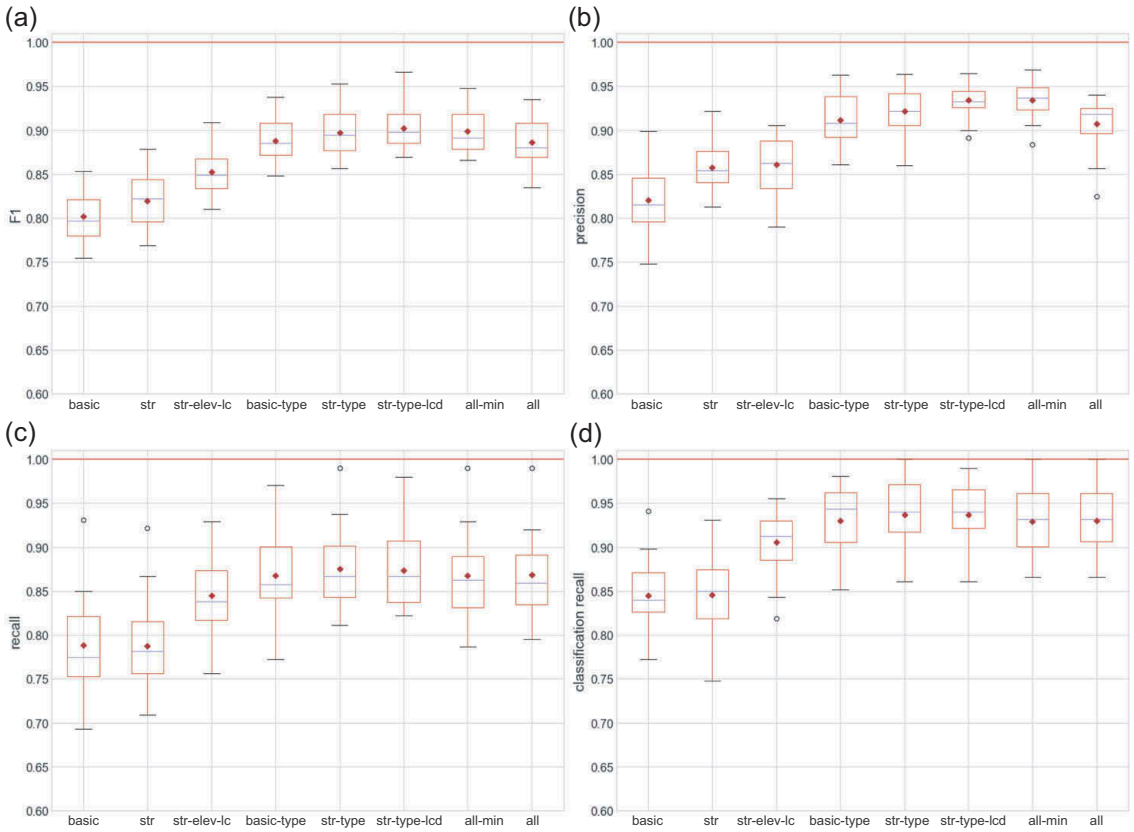


Figure 4. Box plot of medians (blue lines) with interquartile range and means (red diamonds) for: (a) F1 (b) precision (c) overall recall (d) classification recall vs. named combinations of matching features. Feature combinations: *basic*: minimum Levenshtein distance and geographical distance; *str*: all name (string) features and geographical distance; *basic-type*: minimum Levenshtein distance, geographical distance, and encoded feature types; *str-type*: all name (string) features, geographical distance, and encoded feature types; *str-elev-lc*: all name (string) features, geographical distance, elevation, and all land cover features (no feature types); *str-type-lcd*: all name features, geographical distance, encoded feature types, and land cover distance; *all-min*: minimum Levenshtein distance, geographical distance, encoded feature types, elevation, and land cover distance; *all*: all features.

difference between the feature combinations not using encoded feature types (*basic* and *str*) and those that do (the 5 rightmost on the plot), whose worst runs are all better than the median of the former two feature combinations. Somewhere in between we find the *str-elev-lc* combination, which lacks feature type information but uses all other matching features, offering F1 performance above the *basic* and *str* models, but below all models using encoded feature types. By examining the plots for precision and recall, we see that *str-elev-lc*'s performance advantage over *str* is almost entirely due to increased recall.

We obtained strong overall performance using the 5 combinations incorporating feature types as matching features, with no clear advantage of one over the others despite the addition of elevation and land cover features in some combinations (*all-min, all*) but not others (*str-type*). The *str-type-lcd* combination provides the best results over these 20 runs, with the highest median, mean, upper quartile, and lower quartile for F1. In Table 4 we present the mean values for these same runs, alongside the maximum overall recall (*max recall*) obtainable based on the record pairs retained at the candidate

Table 4. Mean values (over 20 runs) of precision, recall, and F1 for named feature combinations. Feature combinations: *basic*: minimum Levenshtein distance and geographical distance; *str*: all name (string) features and geographical distance; *basic-type*: minimum Levenshtein distance, geographical distance, and encoded feature types; *str-type*: all name (string) features, geographical distance, and encoded feature types; *str-elev-lc*: all name (string) features, geographical distance, elevation, and all land cover features (no feature types); *str-type-lcd*: all name features, geographical distance, encoded feature types, and land cover distance; *all-min*: minimum Levenshtein distance, geographical distance, encoded feature types, elevation, and land cover distance; *all*: all features.

feature combination	precision	recall	F1	max recall
basic	0.820	0.788	0.802	
str	0.857	0.788	0.820	
str-elev-lc	0.861	0.845	0.852	
basic-type	0.912	0.868	0.888	0.919
str-type	0.921	0.875	0.897	
str-type-lcd	0.934	0.874	0.902	
all-min	0.934	0.867	0.899	
all	0.907	0.868	0.886	

selection stage. Maximum recall changes with every run since the source records in the test set, and thus their candidate matching target records, also change, but this value is independent of the classification, and thus the mean is constant over all feature combinations.

5.3. Feature-type-balanced test set

In order to get further information about how our machine learning and rule-based matching perform, including how performance varies by feature type and how the machine learning methods respond to increasing amounts of training data, we ran tests using a fixed, feature-type balanced test set (as described in section 4.3.1). The results of these experiments are presented in Figure 5. Figure 5(a) shows the F1 performance on 5 representative feature types for a selection of matching strategies (3 rule-based and 4 machine learning based methods, prefixed by 'rf-' for random forests). This breakdown by type and matching method shows a generally more balanced performance across feature types for the machine learning strategies, particularly those that make use of encoded types (*basic-type, str-type, str-type-lcd*). It is also clear that much of the F1 performance gain of the higher performing strategies comes from doing much better on the worst performing type, streams (STM), while retaining strong performance (F1 >0.8) on the other types. Finally, the rule-based method that considers types via type-specific thresholds (*name-custom-threshold*) is doing better on all the plotted types than the machine learning method which does not consider types (*str*).

Figure 5(b) shows the F1 performance of our machine learning pipeline on the feature-type balanced test set as we increase the size of the training set by incrementally adding 40 randomly chosen source records in a step-wise fashion. Here, the 3 better performing feature combinations show continued improvement with increasing training data, but the *basic* model does not and instead seems to plateau when the training

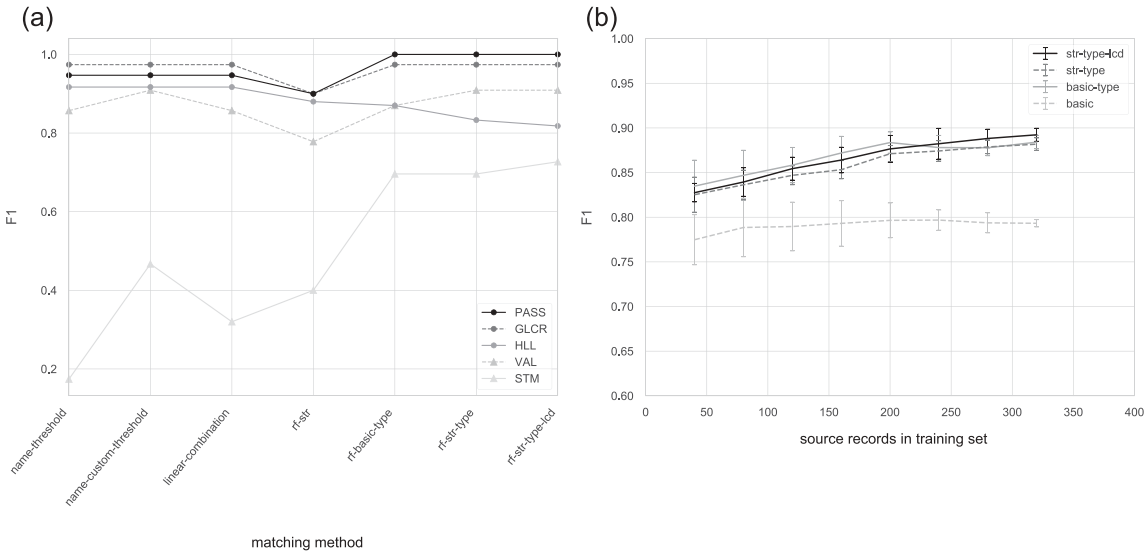


Figure 5. F1 performance according to (a) the matching strategy used (3 rule-based from the left and 4 machine learning based methods from the right, prefixed by *rf-*) broken down by feature type and (b) the number of source records used in the machine learning training pipeline, showing the mean and standard deviation over 10 runs using incrementally more randomly chosen records. Feature combinations plotted: *basic*: minimum Levenshtein distance and geographical distance; *basic-type*: minimum Levenshtein distance, geographical distance, and encoded feature types; *str-type*: all name (string) features, geographical distance, and encoded feature types; *str-type-lcd*: all name features, geographical distance, encoded feature types, and land cover distance.

pipeline uses around 200 source records. In general this means there is potential for our random forests to perform even better than they are, were there more annotated data available for training, above using the full set of 320 source records that are not in the feature-type balanced test set.

6. Discussion

Gazetteer matching is an important, real-world problem, where the very large numbers of records involved mean that small differences in precision or recall can have large implications. In this work, we performed cross-gazetteer matching on a set of natural feature records by implementing rule-based methods and machine learning methods using random forests. Our rule-based methods gave good results (Table 3), but our best machine learning models offered an F1 increase of 6% over the best rule-based results. However, rule-based and machine learning performance was similar when considering only record names and locations. Once feature types were incorporated as matching features in random forests, our models all achieved mean F1 values above 0.88.

This importance of gazetteer feature types on matching performance has received surprisingly limited research attention, with most previous work focusing on a very narrow set of types such as POIs or cities (Zheng *et al*. 2010, Martins 2011, Dalvi *et al*. 2014, McKenzie *et al*. 2014). Despite little guidance on how to effectively handle feature types, doing so was crucial for the natural feature records treated in this work. Thanks to one-hot encoding for categorical variables, incorporating feature types into random forests was simple and enabled the random forests to adapt to types, from multiple

type hierarchies, in a data-driven fashion. In contrast, considering types in rule-based processing was less straightforward, requiring semantic knowledge of the types (e.g. for type-specific distance thresholds), potentially manually aligning type hierarchies (as in Hastings 2008, Morana *et al*. 2014), and tailoring decisions to our particular datasets.

Random forests also offered a more balanced performance across feature types compared to rules, arguably as a result of this flexible approach to types. Though to our knowledge no direct comparison of rules and machine learning has been performed on matching geographical data, a similar finding was obtained in a work on deduplicating a dataset of inventors. Indeed, Ventura *et al*. (2015) compared the performance of rule-based methods against supervised learning using random forests and found that random forests offered much more robust performance with respect to data subsets with varying characteristics. In addition to robustness to feature type profiles, our experiments show that random forests perform increasingly well with more training data, suggesting that extra gains in F1 can be obtained through further annotation. Similar experiments were performed by Zheng *et al*. (2010), who varied the size of their training and testing sets together, and found that overall accuracy increased with the dataset size. Despite these advantages of supervised learning over rules, the performance gains came with costs, including greater complexity and more person-hours spent on implementing a random forest pipeline than on rule-based matching. Since simple rules performed well for the subset of data which was relatively easy (exact or near-exact name matches and very short point-to-point distances), deciding on a matching strategy for a different dataset would require careful consideration of these trade-offs.

Returning to the issue of complexity, implementing a realistic machine learning pipeline required carefully thinking about, and experimentally verifying, processing decisions including how to select match candidates, how to prepare the test set, and what ratio of negatives to positives to use. We found no clear methodological consensus in the literature, and even considerable disagreements, such as on the issue of negative-to-positive pair ratios, with some using 1:1 ratios (Zheng *et al*. 2010, Martins 2011), and others up to 30:1 (Sehgal *et al*. 2006). However, our decision to closely mimic a real-world scenario, treating test records as if they were unannotated, influenced many downstream decisions. This meant letting the testing pipeline find all (positive) matches through candidate selection, and not adding our annotated matches to the test set. This in turn meant splitting source records at the very beginning of the pipeline and choosing a candidate selection method likely to return not just hard negatives, but positive matches – as many as possible to maximize recall. We thus selected candidates based on considering multiple matching features at once (similar to Zheng *et al*. 2010, Morana *et al*. 2014), avoiding techniques geared more specifically towards negative selection (Sehgal *et al*. 2006, Martins 2011) and which could potentially limit recall. Finally, choosing how many matching candidates to keep per source record (i.e. k) was a purely experimental decision, optimizing for F1 on the full pipeline. We found that low values of k (in other words, low ratios of negatives to positives, assuming an average positive match count around 1 per record) limited recall and settled on a k of 30, similar to Sehgal *et al*. (2006).

After this closer look at how our methodology compares to existing work, a related question is, how do our results compare with previous published results? From our survey of

Table 5. Data-related aspects of selected papers comparable to the present work.

Authors	Task	Data description	Feature types
Sehgal *et al*. 2006	Gazetteer matching	US & UK authoritative data for Afghanistan	Varied (all)
Hastings 2008	Gazetteer matching	3 datasets covering Lake Tahoe (US)	Administrative, cultural, and water
Smart *et al*. 2010	Gazetteer matching	UK data from: GeoNames, Ordnance Survey, OpenStreetMap, Yahoo! Where on Earth, Wikipedia	Varied (all)
Zheng *et al*. 2010	Deduplication	POIs and yellow page records for Beijing (China)	POIs
Martins 2011	Deduplication	Global, mixed source	Populated places
Gonçalves 2012	Deduplication	Global, mixed source	Populated places
McKenzie *et al*. 2014	Gazetteer matching	Foursquare, Yelp (US)	POIs
Dalvi *et al*. 2014	Deduplication	Facebook Places (US)	POIs

Table 6. Method-related aspects of selected papers comparable to the present work.

					Best performance		
Authors	Approach	Method summary	Annotated positive pairs	neg:pos ratio	p	r	f1
Sehgal *et al*. 2006	Machine learning	Logistic regression, voted perceptron, SVM	2,006	30:1	.96	.92	.94
Hastings 2008	Rule-based	Consider names, footprints, feature types	252	N/A	.89	.22	.35
Smart *et al*. 2010	Rule-based	Consider names and locations	16	N/A	1.0	.44	.61
Zheng *et al*. 2010	Machine learning	Decision tree with bootstrap aggregating	800	1:1	.89	.87	.88
Martins 2011	Machine learning	SVM, alternating decision tree	1,927	1:1	.99	.98	.98
Gonçalves 2012	Machine learning	SVM, alternating decision tree, random forest	4,401	1:1	.97	.97	.97
McKenzie *et al*. 2014	Regression	Weighted multi-attribute model	100	N/A	accuracy = .97		
Dalvi *et al*. 2014	Machine learning	Unsupervised language model using local context	4,000	7:2	.90	.90	.90

the literature, it is clear that the variety of datasets and methods used make comparison difficult. Nonetheless, we have compiled a list of the most comparable papers and present a structured summary of these, with data-related aspects in Table 5, and method-related aspects, including best reported performance, in Table 6. We list 'task' as a data-related aspect since the key difference between a deduplication task and a (cross-)gazetteer matching task lies in whether data are already structured according to a single schema and single feature type hierarchy. With a single type hierarchy, a range of additional matching features can be used, for instance type equality or the hierarchical distance between types (Zheng *et al*. 2010, Martins 2011).

Visible is the predominance of datasets featuring POIs and populated places, which could arguably simplify matching due to low feature type diversity and, especially for POIs, a predictable spatial granularity for the records. POIs however present their own challenges, including potentially very large dataset sizes (Zheng *et al*. 2010, Dalvi *et al*. 2014) and particular naming patterns, which Zheng *et al*. (2010) tackle using custom token-based name similarity metrics, and Dalvi *et al*. (2014) using an innovative unsupervised language modeling approach. These two works achieve results similar to ours, with F1 values reaching 0.88 (Zheng *et al*. 2010) and 0.90 (Dalvi *et al*. 2014). McKenzie

et al. (2014) also report on a dataset of POIs, developing a regression approach with a wide set of matching features, but their evaluation is mainly performed on an artificial test set where 100 exact matches were created. In this rather unrealistic testing scenario, they report an accuracy of 0.97, but mention performance decreases when random pairs are selected from data, reporting F-scores of 0.35 and 0.32. In Martins (2011) and a follow-up work (Gonçalves 2012), very high deduplication performance is reported, with F1 values reaching 0.98 and 0.97, respectively. However, the authors note that their dataset of coarse-grained records (populated places) may not be particularly challenging, with F1 values reaching 0.97 when using only matching features based on name similarity (Martins 2011).

Two other works (Hastings 2008, Smart *et al*. 2010) are broadly similar to each other in that they use rule-based processing and datasets with diverse feature types, but don't rigorously evaluate their approaches. Smart *et al*. (2010) manually examine only a very small number of individual cases (n = 16), achieving perfect precision, but a nominal recall of only 0.44, and Hastings (2008) present a descriptive evaluation, for which we calculated precision to have been 0.89, and recall only 0.22. In contrast, Sehgal *et al*. (2006) perform a rigorous quantitative evaluation of their machine learning based matching, reporting a range of values for precision, recall, F1, and accuracy, including how these change as a function of the ratio of negatives-to-positives. We perform comparable tests on setting a value of *k* (Figure 3), and complement this more extensive evaluation by performing multiple runs to test robustness to both individual training data sets and different training set sizes.

A final important point about our ability to compare our results to previous work is that it is at times unclear how matches are found for test records. If a single static collection of positive and negative record pairs is used to train and test a classifier, for example via cross-validation, then it is likely that the full set of annotated positives automatically finds it way into each testing fold (Figure 6). This would optimistically bias performance compared to a scenario in which matches for test records must be found via candidate selection, and from a large number of such candidate records, as is the case in our pipeline. In addition, adding random negatives into a test set will

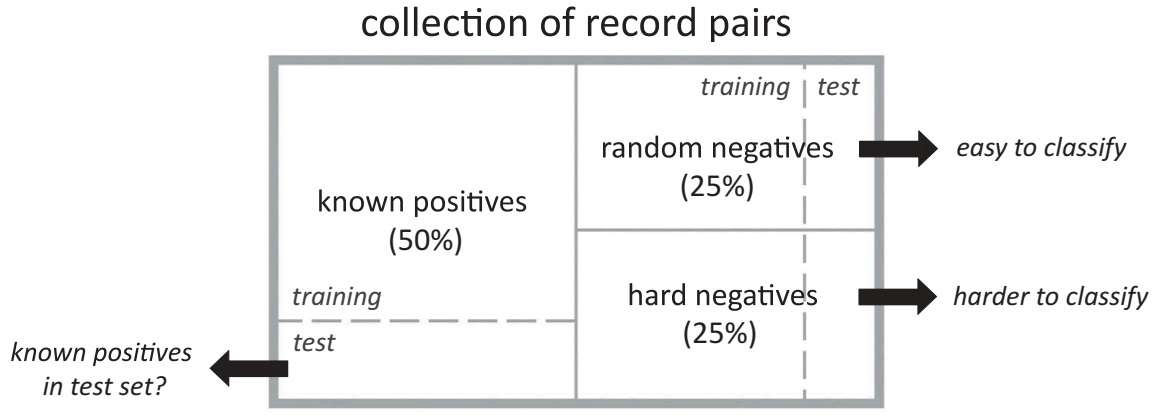


Figure 6. Potential test set composition when using cross-validation with a collection of record pairs built with a 1:1 negative to positive ratio and where 50% of negatives are chosen randomly.

optimistically bias accuracy, since these record pairs will overwhelmingly be highly dissimilar, and thus easy to classify as non-matches (Figure 6). Whether random negatives have any effect on precision and recall is unclear, but we opted against using them in our training pipeline since we already had quite varied training data, thanks in part to using a high value of *k*. Furthermore, random forests rely on random samples of the training data for each tree and thus naturally include variations from different samples of the distribution.

7. Conclusion

In this work, we tackled the problem of matching natural feature records across two gazetteers. We showed that good performance could be obtained using relatively simple rules, but that machine learning using random forests offered not only better performance, but greater flexibility, obviating the need to manually align feature types and tune thresholds. Random forests also offered a more balanced performance across feature types and the potential for even better performance by increasing the amount of training data through further annotation. We emphasize that creating a training dataset in order to implement a machine learning solution was both more straightforward than handcrafting rules for gazetteer matching and more easily generalizable. With random forests, the biggest performance increase over basic models using string similarity and geographical distance came from incorporating feature types as matching features, after which all tested models performed similarly well. However, all classifiers which included some representation of feature type, including by using land cover classes and elevation instead of gazetteer feature types, outperformed simpler string and geographical distance based classifications.

Although our results were obtained on the specific case of matching records between a Swiss national gazetteer and GeoNames, they have more general implications, particularly in the context of growing interest in spatial data science. We make the following recommendations for future work in this area:

- Gazetteer matching is influenced by feature types and therefore any processing decisions related to these feature types for matching should be explicitly described.
- Training and evaluation datasets should be carefully designed so as not to make classification problems unrealistically straightforward or difficult (e.g. nearby toponyms with similar names which are not matches should be explicitly included in test data). Future work should also consider the impact of candidate selection on overall performance.
- Since gazetteer data are often snapshots, and may also not be freely available, making at a minimum annotated data available will both increase the potential for reproducibility, and allow other researchers to understand the properties of the snapshots investigated. Use of shareable markdown (e.g. R-Markdown, Juypter notebooks) will further increase reproducibility and make all stages of processing more transparent.

Notes

1. The term 'feature' is both widely used in the GIScience community to refer to geographical entities and in the machine learning community to refer to properties of the data used to

train models. To avoid potential ambiguity, we refer to geographical features as 'entities' in a real-world context or 'records' in a gazetteer context, and we use 'features' to refer to 'matching features' in the machine learning sense. We however maintain the use of the widely used two-word expressions 'feature type' to refer to the catalogued type of geographical entities and 'natural features' to refer to the subset of geographical entities which are not human-made.

2. https://shop.swisstopo.admin.ch/en/products/landscape/names3D.
3. CH.zip from http://download.geonames.org/export/dump/.
4. https://pandas.pydata.org/.
5. https://github.com/eacheson/machine-learning-gazetteer-matching.
6. https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/geostat/geodaten-bundesstatistik/boden-nutzung-bedeckung-eignung/arealstatistik-schweiz/bodenbedeckung.html.

Acknowledgments

RSP gratefully acknowledges support from Swiss National Science Foundation Project EVA (166788).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung [166788].

Notes on contributors

Elise Acheson is a PhD student in the Geocomputation Unit at the University of Zurich. Her work focuses on automatically extracting geographical information from various textual sources.

Michele Volpi is a Senior Data Scientist at the Swiss Data Science Center, a joint venture between ETH Zurich and EPFL aiming at accelerating the adoption of data science within the ETH Domain and the Swiss academic community at large. His main research activities are at the interface of computer vision, machine learning, and deep learning, focusing on the extraction of information from aerial and satellite imagery and from geospatial and environmental data in general.

Ross S. Purves is a Professor at the Department of Geography of the University of Zurich. His research interests focus on how we can answer and explore societally relevant geographic questions paying attention to vagueness and uncertainty, often using unstructured data in the form of text as a primary source.

References

Acheson, E., *et al*., 2017b. Gazetteer matching for natural features in switzerland. *In*: Christopher B. Jones and Ross S. Purves, eds. *Proceedings of the 11th Workshop on Geographic Information Retrieval, GIR'17* New York, NY, USA: ACM, 11:1–11: 2.

Acheson, E., De Sabbata, S., and Purves, R.S., 2017a. A quantitative analysis of global gazetteers: patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64, 309–320. doi:10.1016/j.compenvurbsys.2017.03.007

Adams, B., McKenzie, G., and Gahegan, M., 2015. Frankenplace: interactive thematic mapping for Ad Hoc exploratory search. *In*: Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi, eds. *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 12–22.

Ahlers, D., 2013. Assessment of the accuracy of GeoNames gazetteer data. *In*: Christopher B. Jones and Ross S. Purves, eds. *Proceedings of the 7th Workshop on Geographic Information Retrieval, GIR '13*, New York, NY, USA: ACM, 74–81.

Berman, M.L., Mostern, R., and Southall, H., eds.. 2016. *Placing names: enriching and integrating gazetteers*. Bloomington, Indiana: Indiana University Press.

Brauner, D.F., Casanova, M.A., and Milidiú, R.L., 2007. Towards gazetteer integration through an instance-based thesauri mapping approach. *In*: C.A.D. Jr and A.M.V. Monteiro, eds. *Advances in geoinformatics*. Campos do Jordão (SP), Brazil: Springer Berlin Heidelberg, 235–245.

Breiman, L., 2001. Random forests. *Machine Learning*, 45 (1), 5–32. doi:10.1023/A:1010933404324

Brunner, T.J. and Purves, R.S., 2008. Spatial autocorrelation and toponym ambiguity. *In*: Chris Jones and Ross Purves, eds. *Proceedings of the 5th Workshop on Geographic Information Retrieval, GIR '08* New York, NY, USA: ACM, 25–26.

Christen, P., 2012. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24 (9), 1537–1555. doi:10.1109/TKDE.2011.127

Cooper, D. and Gregory, I.N., 2011. Mapping the English lake district: a literary GIS. *Transactions of the Institute of British Geographers*, 36 (1), 89–108. doi:10.1111/tran.2010.36.issue-1

Costa, G., 2011. Data de-duplication: a review. *In*: J. Kacprzyk, *et al*., eds. *Learning structure and schemas from documents*, Vol. 375. Berlin, Heidelberg: Springer Berlin Heidelberg, 385–412.

Dalvi, N., *et al*., 2014. Deduplicating a places database. *In*: Chin-Wan Chung, Andrei Broder, Kyuseok Shim, and Torsten Suel, eds. *Proceedings of the 23rd International Conference on World Wide* Web, WWW '14, New York, NY, USA: ACM, 409–418.

Elmagarmid, A.K., Ipeirotis, P.G., and Verykios, V.S., 2007. Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 19 (1), 1–16. doi:10.1109/TKDE.2007.250581

Fu, G., Jones, C.B., and Abdelmoty, A.I., 2005. Building a geographical ontology for intelligent spatial search on the web. *In*: M.H. Hamza, ed. *Proceedings of IASTED International Conference on Databases and Applications (DBA-2005)*, February. Innsbruck, Austria: ACTA Press, 167–172.

Gao, S., *et al*. 2017. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems*, 61 (Part B), 172–186. doi:10.1016/j.compenvurbsys.2014.02.004

Gelernter, J., *et al*., 2013. Automatic gazetteer enrichment with user-geocoded data. *In*: Dieter Pfoser and Agnès Voisard, eds. *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Inf*ormation, GEOCROWD '13, New York, NY, USA: ACM, 87–94.

Gonçalves, N.F.A., 2012. Gazetteer record linkage. Master's thesis. Lisbon: Instituto Superior Técnico. doi:10.1094/PDIS-11-11-0999-PDN

Hastings, J.T., 2008. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, 22 (10), 1109–1127. doi:10.1080/13658810701851453

Hill, L.L., 2006. *Georeferencing: the geographic associations of information*. Cambridge, MA: The MIT Press.

Janowicz, K. and Keßler, C., 2008. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 22 (10), 1129–1157. doi:10.1080/13658810701851461

Keßler, C., *et al*. 2009. Bottom-up gazetteers: learning from the implicit semantics of geotags. *In*: K. Janowicz, M. Raubal, and S. Levashkin, eds. *GeoSpatial Semantics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 83–102.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii nauk SSSR*, 163 (4), 845–848.

Lieberman, M.D., Samet, H., and Sankaranayananan, J., 2010. Geotagging: using proximity, sibling, and prominence clues to understand comma groups. *In*: Ross Purves, Paul Clough, and Chris

Jones, ed. *Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR '10*, New York, NY, USA: ACM, 6: 1–6: 8.

Martins, B., 2011. A supervised machine learning approach for duplicate detection over gazetteer records. *In*: Christophe Claramunt, Sergei Levashkin, and Michela Bertolotto, eds. *GeoSpatial semantics*, Lecture notes in computer science, May. Berlin, Heidelberg: Springer, 34–51.

McKenzie, G., Janowicz, K., and Adams, B., 2014. A weighted multi-attribute method for matching user-generated points of interest. *Cartography and Geographic Information Science*, 41 (2), 125–137. doi:10.1080/15230406.2014.880327

Morana, A., *et al*., 2014. GeoBench: a geospatial integration tool for building a spatial entity matching benchmark. *In*: Yan Huang, Markus Schneider, Michael Gertz, John Krumm, and Jagan Sankaranarayanan, eds. *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, New York, NY, USA: ACM, 533–536.

Murphy, A.H., 1996. The finley affair: a signal event in the history of forecast verification. *Weather and Forecasting*, 11 (1), 3–20. doi:10.1175/1520-0434(1996)011<0003:TFAASE>2.0.CO;2

Olteanu, A., Musti'ere, S., and Ruas, A., 2006. Matching imperfect spatial data. *In*: M. Caetano and M. Painho, eds. *Proceedings of 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, July. Lisbon, Portugal, 7–9.

Pedregosa, F., *et al*. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

Popescu, A., Grefenstette, G., and Mo¨Ellic, P.A., 2008. Gazetiki: automatic creation of a geographical gazetteer. *In*: Ronald Larsen, Andreas Paepcke, José Borbinha, and Mor Naaman, eds. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '08*, New York, NY, USA: ACM, 85–93.

Purves, R.S., *et al*. 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science*, 21 (7), 717–745. doi:10.1080/13658810601169840

Santos, R., *et al*., 2018. Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, Taylor & Francis, 32 (2), 324–348. doi:10.1080/13658816.2017.1390119.

Sehgal, V., Getoor, L., and Viechnicki, P.D., 2006. Entity resolution in geospatial data integration. *In*: Rolf A. de By, and Silvia Nittel, eds. *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems, GIS '06*, New York, NY, USA: ACM, 83–90.

Smart, P.D., Jones, C.B., and Twaroch, F.A., 2010. Multi-source toponym data integration and mediation for a meta-gazetteer service. *In*: Sara Irina Fabrikant, Tumasch Reichenbacher, Marc van Kreveld, and Christoph Schlieder, eds. *Geographic information science*, Lecture notes in computer science, September, Berlin, Heidelberg: Springer, 234–248.

Smith, B. and Mark, D., 2003. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B (Planning and Design)*, 30 (3), 411–427. doi:10.1068/b12821

Ventura, S.L., Nugent, R., and Fuchs, E.R.H., 2015. Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy*, 44 (9), 1672–1701. doi:10.1016/j.respol.2014.12.010

Vincenty, T., 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 23 (176), 88–93. doi:10.1179/sre.1975.23.176.88

Walter, V. and Fritsch, D., 1999. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science*, 13 (5), 445–473. doi:10.1080/136588199241157

Zheng, Y., *et al*., 2010. Detecting nearly duplicated records in location datasets. *In*: Divyakant Agrawal, Pusheng Zhang, Amr El Abbadi, and Mohamed Mokbel, eds. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10*, New York, NY, USA: ACM, 137–143.

Zhu, R., *et al*., 2016. Spatial signatures for geographic feature types: examining gazetteer ontologies using spatial statistics. *Transactions in GIS*. doi:10.1111/tgis.2016.20.issue-3

Paper III

Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach

Summary

This paper presents and uses a methodology to gather and compare landscape descriptions from three different textual sources (free lists from in-person interviews, Flickr photo tags, and hiking blogs). As part of the methodology, geographically-focused areal footprints are generated from the hiking blogs for each study site, using a custom-built text-to-space pipeline, and these footprints are used to spatially query Flickr data. With textual data sources linked in geographical space, an analysis is carried out to examine how the study site, landscape type, and data source type relate to the information content in our data.

Contribution of the PhD candidate

I drafted parts of the manuscript (co-authored 'Methodology' and 'Results and interpretation' and regularly edited the rest of the manuscript). I wrote the Python analysis code for the text-to-space pipeline, as well as for the cosine similarity comparisons and Mann-Whitney-U tests.

Citation

Wartmann, F. M., Acheson, E., and Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach. *International Journal of Geographical Information Science*, 32(8):1572–1592. DOI:10.1080/13658816.2018.1445257

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2018
https://doi.org/10.1080/13658816.2018.1445257

Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach

Flurina M. Wartmann [a], Elise Acheson[a] and Ross S. Purves[a,b]

[a]Geography Department, University of Zurich, Zurich, Switzerland; [b]Research Priority Programme on Language and Space, University of Zurich, Zurich, Switzerland

ABSTRACT

How do people perceive landscapes? What elements of the landscape do they identify as characteristic of a landscape? And how can we arrive at descriptions, and ultimately representations that better reflect people's notions of landscapes? In this study, we collected landscape descriptions from five landscape types at 10 study sites in Switzerland. For each site, we collected data from three sources: free lists with participants, hiking blogs, and Flickr tags. Free lists were obtained through on-site interviews with visitors, hiking blogs were gathered by focused crawling of web content, and Flickr tags were selected based on spatial footprints obtained from the hiking blogs. We quantitatively compared landscape descriptions between data sources and landscape types using cosine similarity. We found that descriptions from the same data source were significantly more similar, irrespective of landscape type. Descriptions from the same landscape type were more similar, but only within the same data source. Through a qualitative analysis of different aspects of landscape in our content, we found that each data source offered a different distribution of landscape aspects. For example, while Flickr tags contained high proportions of toponyms, they contained little content relating to sense of place. In contrast, hiking blogs contained more information about sense of place. Our approach combining these varied textual sources thus offers a more holistic view on landscapes. This study constitutes a step toward extracting semantically rich descriptions of landscapes from a variety of sources and using this information to distinguish different landscapes, with potential applications for landscape monitoring and management.

ARTICLE HISTORY
Received 4 October 2017
Accepted 21 February 2018

KEYWORDS
Place descriptions; landscape character assessments; user-generated content; geographic information retrieval; text processing

1. Introduction

How do people perceive landscapes? What elements of the landscape do they identify and pick out as being characteristic? And how can we arrive at descriptions, and ultimately computational representations that better reflect people's notions of landscapes? Making a methodological contribution to answering these questions lies at the heart of this article's aims. In particular, we wish to make a contribution to research methods that relate to landscape policy. An important starting point is thus the European Landscape Convention (Council of Europe 2000), which defines landscape

CONTACT Flurina M. Wartmann flurina.wartmann@ed.ac.uk

as being 'an area, as perceived by people, whose character is the result of the action and interaction of natural and/or human factors'. Central to this definition is the notion of landscape as being something which is perceived and thus, presumably, characterizations of landscape should consider ways in which they are described by people (Scott 2002, 2003).

Methods to characterize and assess landscapes have a long history in landscape research (Daniel and Vining 1983, Zube 1984, Brabyn 1996). The importance of such methods has increased, since less tangible benefits derived by humans through landscapes form a cornerstone of the ecosystem services (ES) framework. ES aim to quantify the benefits which humans derive, either directly or indirectly, from ecosystems (Costanza *et al*. 1997, MA 2005) and can be broadly divided into four classes: provisioning (e.g. the provision of fresh water or biomass); regulating (e.g. the ability of a flood plain to reduce flooding downstream); supporting (e.g. the redistribution of seeds by birds); and cultural (e.g. the recreational or spiritual meaning attached to an ecosystem) (MA 2005). Perhaps unsurprisingly, efforts to map ES often turn to spatially continuous data and apply traditional GIS analytical techniques to locating and quantifying ES (de Groot *et al*. 2010). However, although this approach may function well in modelling ES with a relatively direct relationship to biophysical properties, such as estimates of aboveground biomass through tree cover and derived volume (Dong *et al*. 2003) or the number of visitors to a region (Nahuelhual *et al*. 2013), it is less well suited to capturing and representing many cultural values. Indeed, in a call strikingly similar to the social critique of GIS advanced by Pickles (1995), Kirchhoff (2012) concluded that: '[.] pivotal cultural values attaching to the natural/cultivated environment cannot be integrated into the ES framework, and should not be called cultural ES'.

We do not propose to revisit these debates here, but rather suggest that this problem has to do with not only what we try to model, but also how we go about doing so. If we wish to model culturally meaningful properties, then an appropriate starting point is not a continuous representation of space such as a land cover/land use map, but rather the landscape features with which cultural meanings are associated (Mark and Turk 2003, Kirchhoff 2012). Since such features are not readily embedded in spatially explicit, continuous representations (Smith and Mark 2003), we take as our starting point not spatial data, but spatially grounded language. This approach can be seen to overlap with recent efforts to use crowdsourced information to characterize different aspects of landscapes (e.g. van Zanten *et al*. 2016) with the underlying assumptions that, firstly, such efforts potentially allow us to characterize large areas efficiently and, secondly, that data gathered through crowdsourcing are representative of the underlying properties of landscape in which we are interested.

In this article, we explore methods to collect and compare landscape descriptions obtained from the public through three approaches. Our aim is to analyze to what extent different methodological approaches, and consequently, different data sources, result in different descriptions of landscapes, and in doing so demonstrate the potential of combining complementary approaches to characterize landscapes bottom-up. To address this aim, we specify two research questions:

- RQ1: How can empirical *in situ* methods be combined with data-driven approaches to collect landscape descriptions from the public?
- RQ2: How do landscape descriptions from different data sources differ?

We chose 10 study sites in Switzerland to empirically investigate these research questions. Using a triangulation of methods based on *in situ* free listings with visitors, full text descriptions mined from the web, and georeferenced image tags from the photo-sharing platform Flickr, we compared landscape descriptions with respect to data sources and landscape types. Our hypothesis is that landscape types reflect differences perceived by people, which are reflected in differences in the textual descriptions of these landscapes.

The remainder of this article is organized as follows: in Section 2 we introduce related work, before describing the study sites and the methods used to collect and analyze landscape descriptions in Section 3. In Section 4 we present results from comparisons of data sources and landscape types, interpreting these with respect to landscape characterization and typologies in the Swiss context. In Section 5 we discuss our results in relation to other work and our research questions, before concluding the article in Section 6.

2. Related work

Since our aim in this article is to characterize landscapes using multiple data sources in ways which are useful in both science and policy, we first briefly establish the need for improved methods to characterize landscapes, highlighting some of the key challenges. We then give a brief overview of the methodological underpinnings of the three approaches we chose to apply to the problem, identifying properties of both the data sources and the methodological tools necessary for their analysis in the context of spatially situated landscape descriptions.

2.1. *Challenges in landscape characterization*

Any characterization of landscape must first deal with two related fundamental questions. Firstly, should landscape be essentially treated as a continuous field, or as a set of identifiable objects to which properties are attached (Mark *et al*. 2011)? And secondly, can landscape properties be objectively extracted from spatial data or are they the product of individual perception (Warnock and Griffiths 2015)? We suggest these dichotomies are linked to the definition of landscape underpinning its characterization, and to the methods applied. Approaches treating landscape as a set of features that are identified and named with generic terms (e.g. stream, hedge, meadow) pay particular attention to the importance of language in structuring our experience of landscape (Johnson and Hunn 2010, Mark *et al*. 2011), but these terms are seldom used as the basis for mapping (Wartmann and Purves 2017). Approaches conceptualizing landscape as an objective, continuous field often focus on extracting landscape units from available data, based on shared biophysical and morphological properties, for example in the form of terrain attributes or land cover classes using supervised or unsupervised approaches (Bunce *et al*. 1996, Gerçek *et al*. 2011, Niesterowicz *et al*. 2016). However, such approaches typically either apply existing expert classifications of input data, as is the case when land cover data are used (Mücher *et al*. 2010, Comber 2013), or attach relatively simple semantics to extracted patterns (Iwahashi and Pike 2007).

Missing in such approaches is a direct link to how landscape is subjectively perceived, despite this being recognized as a requirement in legislation, for example in the European Landscape Convention (Council of Europe 2000). This gap has led to the development of approaches that attempt to characterize landscape based not only on its biophysical properties, but also more directly based on human perception (Warnock and Griffiths 2015).

In Switzerland, where we conducted our case study, a landscape typology exists with 38 landscape types, which are modelled using a range of mostly biophysical criteria such as geology, geomorphology, climate, topography, and land use (ARE 2011a). However, perceptual aspects are also recognized, although these are based on expert assessments (ARE 2011b), in which the views of outsiders dominate more locally grounded ones (Butler 2016). The Swiss Landscape Monitoring framework 'LABES' goes further, including both biophysical indicators and those related to cultural perception of landscapes, such as 'distinctiveness', 'authenticity', 'fascination', and 'perceived landscape beauty' (Kienast *et al*. 2015). Assessing these indicators involved sending out written questionnaires to Swiss households, with a total of 2800 questionnaires returned for analysis (Kienast *et al*. 2015). Written questionnaires, however, incur high costs, typically with relatively low response rates allowing analysis only at coarse spatial granularities, and for spatial extents that are defined by administrative boundaries, which may be at odds with how people perceive and experience landscapes.

In the United Kingdom, a long tradition of including perceptual and aesthetic aspects for landscape management and planning exists in the form of Landscape Character Assessments (LCAs) and associated approaches (Swanwick 2002, Natural England 2014, Sarlöv Herlin 2016). Landscape character is considered as the 'distinct and recognizable pattern of elements that occur consistently in a certain type of landscape. […] Character is what makes landscapes distinctive and creates a particular sense of place in a locality' (Swanwick 2004, p.111).The guidelines developed by the Countryside Agency and the Scottish Natural Heritage, and other similar initiatives, have also been adapted in other countries (Jessel 2006, Caspersen 2009, Van Eetvelde and Antrop 2009a, 2009b). Typically, experts assess both environmental and cultural aspects and produce outputs including textual descriptions and sketches describing relatively homogeneous regions. The views of the public may be included through a variety of empirical methods including questionnaires, group workshops, and participatory mapping (Swanwick *et al*. 2002, Caspersen 2009). Key to the resulting products are not only maps of bounded landscape units, but rich textual descriptions associated with these regions. These approaches come closer to incorporating ways in which landscapes are perceived, but are time-consuming and often still top-down, with limited incorporation of public perception (Butler 2016).

The emergence of large volumes of user-generated content, or volunteered geographic information, presents an opportunity to capture greater volumes of data with respect to landscape perception, especially in the form of images and associated descriptions. Indeed, georeferenced images have been used as proxies of landscape preference at both regional (Tenerelli *et al*. 2016, Yoshimura and Hiura 2017) and continental scales (van Zanten *et al*. 2016). The potential value of image descriptions has also been recognized in exploring landscapes (Dunkel 2015) and extracting information related to cultural ES (Richards and Friess 2015, Guerrero *et al*. 2016, Figueroa-Alfaro

and Tang 2017, Oteros-Rozas *et al*. 2017). User-generated content can at least start to bridge the data gap between expert and bottom-up perceptions of landscape, and provide more spatially extensive data, at lower costs. However, it remains to be shown to what extent different data sources and approaches can complement one another.

2.2. *Bottom-up approaches to collect data about landscape characteristics*

In this article, we took three approaches to characterizing landscapes. The first is based on elicitation through free-listing exercises. Although inferior to more detailed ethnographic methods (Johnson and Hunn 2010, Mark *et al*. 2011), free-listing tasks in classroom settings have been shown to elicit landscape categories which appear to converge across languages in European and US settings (Mark *et al*. 1999, Giannakopoulou *et al*. 2013). Adding more context to the free-listing task, either by using video stimuli (Williams *et al*. 2012) or interviewing participants outdoors in different landscape settings (Wartmann 2015), has shown that participants name locally relevant landscape features, using cognitive associations to recall individual terms. However, though free listing is relatively straightforward, conducting *in situ* experiments is still time-consuming.

Our second method is based on user-generated content in the form of unstructured text. Recent work on the analysis of digitized text corpora has started to reveal the potential richness of unstructured text as a means for collecting information about landscapes. For instance, Derungs and Purves (2013) georeferenced and extracted terms related to natural landscape features from a Swiss alpine mountaineering corpus. They used these features to compare landscape descriptions across space through spatially weighted term vectors of natural landscape features, generating so-called 'spatial folksonomies' (Derungs and Purves 2016). These approaches characterized landscape, at least at the level of landscape features, bottom-up, but retained a focus on a continuous, grid-based representation. More generally, there is an ongoing recognition of the potential of such unstructured text as a source of geographic information, with the important proviso that relating text to specific locations remains a challenging task (Gregory and Hardie 2011, Wang and Stewart 2015).

Our third and final approach uses image tags from user-generated content. As image tags emerged as a relevant textual data source, one important question was the extent to which it was comparable with previous empirical work. Edwardes and Purves (2007) showed that lists of terms extracted from an image sharing website broadly followed similar patterns as previous work based around free lists (Mark *et al*. 1999), while Rorissa (2008) demonstrated that tags associated with individual images often took the form of cognitive basic levels which also appear to emerge from free-listing experiments (Tversky and Hemenway 1983). However, in contrast to free-listing experiments, instances (as opposed to types) are acknowledged as being very common ways of tagging individual images, with for example 25 percent of tags reported to take the form of toponyms (Hollenstein and Purves 2010). Since a portion of Flickr images is also georeferenced, it is possible to both define regions based on co-occurring Flickr tags (Grothe and Schaab 2009, Hollenstein and Purves 2010), and to characterize regions based around tag occurrence (Rattenbury and Naaman 2009, Gschwend and Purves 2012, Dunkel 2015). However, a major shortcoming of such approaches remains the

nature of the vocabulary used in tagging, and the difficulties of defining meaningful units with which to associate descriptions.

Based on the challenges associated with each of these approaches, we integrated them into a single methodology that we now outline.

3. Methodology

3.1. *Overview*

In support of our overall aim to explore approaches to characterizing landscapes through language, we collected landscape descriptions from 10 study sites using three different approaches, each approach resulting in a particular textual data source for further analysis. Thus, interviewing participants in the field resulted in free lists, focused web crawling resulted in a corpus of hiking blogs relating to each study site, and querying georeferenced Flickr photos resulted in lists of image tags for each location. Since we wished to compare descriptions not only with respect to the data source, but also with respect to different landscape types, we selected five landscape types that reflect the diversity of landscapes at the intersection between cultural and natural landscapes in Switzerland: mountain, moor, lake, river, and hill landscapes. These landscape types are informed by the formal Swiss landscape typology (ARE 2011b), except for lake landscapes, which are not a recognized landscape category in the formal typology, but which we included based on the importance of water bodies in landscape preference (Pitt 1989). For each of the five landscape types, we selected two study sites based on the criteria of accessibility through hiking paths and public transport, and high visitor numbers, for a total of 10 study sites (Table 1).

With three approaches to gathering landscape descriptions and 10 study sites, our final collection to analyze and compare consisted of 30 'documents', one for each combination of study site (Table 1) and data source (free lists, hiking blogs, Flickr tags). We borrow here the term 'document', commonly used in the information retrieval community to mean a data object (prototypically unstructured text) that can be processed, indexed, and queried. Each data collection approach is described in more detail in Section 3.2, and an overview is presented in Figure 1. After collecting all our textual descriptions, we extracted and categorized terms according to which of a set of landscape aspects they best pertained to, described in Section 3.3. We then compared the distribution of these

Table 1. Study sites and their landscape types in the Swiss landscape typology.

Location	Landscape type
Oeschinensee	Limestone mountain landscape of the Alps
Seealpsee	Limestone mountain landscape of the Alps
Thurauen, River Thur	River landscape
Bremgarten, River Reuss	River landscape
Robenhuserriet, Pfäffikon	Moor-influenced landscape
Ägeriried, Rothenthurm	Moor-influenced landscape
Ufschötti, Lake Lucerne	Urban landscape [with lake]
Zürichhorn, Lake Zurich	Urban landscape [with lake]
Hochwacht, Lägern	Landscape of hills of the Central Plateau with a focus on forage production
Hochwacht, Pfannenstiel	Landscape of hills of the Central Plateau with a focus on agricultural production

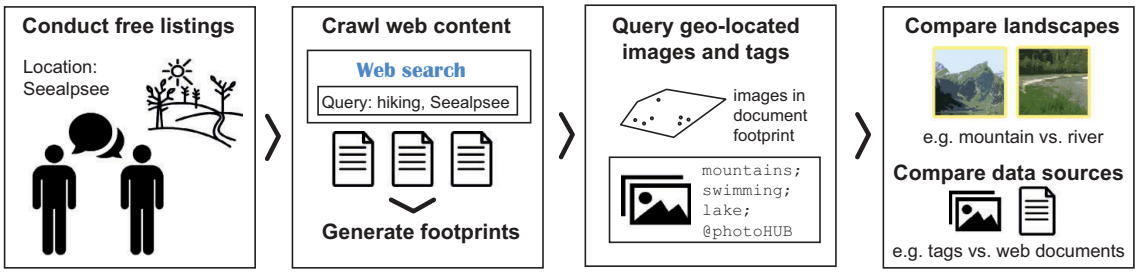


Figure 1. Overview of the methodological sequence applied to extract and compare landscape descriptions.

landscape aspects across our three data sources. Finally, we used the landscape aspect categories to filter terms before calculating similarity measures between documents. We calculate the statistical significance of the similarity measure across sets of documents to compare both data sources and landscape types (Section 3.4).

3.2. *Collecting landscape terms*

At each study site, we first conducted free-listing tasks with 30 visitors. A detailed description of our free-listing approach can be found elsewhere (Wartmann and Purves, in press). For the sake of understanding, we here include a brief overview of the method. We selected visitors pragmatically, while attempting to achieve a balance across different age groups and gender. At the interview locations, we approached visitors and asked if they were willing to take part in a study. If they agreed, we conducted a free-listing task using the elicitation statement in Swiss German: 'Was hätts für Sie i dere Landschaft?', which can literally be translated as 'What is there for you in this landscape'? Participants were instructed to list whatever came to their mind, and that there were no right or wrong answers. If participants paused during their lists, they were prompted once as to whether they wanted to continue. Participants were also asked a basic set of demographic questions. In total, we thus interviewed 300 participants (155 men, 145 women). Of the participants, 60 were aged 34 and under, 106 were between 35 and 54 years old, while 118 were 55 and older (16 participants did not share their age). The first author of this article, a native speaker of Swiss German, transcribed all answers during the interviews as lists of terms. These *free lists* formed our first set of documents.

For each study site, we then created a small web-crawled corpus of full text descriptions about landscapes, consisting mainly of hiking blogs. We used the open-source tool BootCaT (Baroni and Bernardini 2004) to create these web corpora by using a query consisting of toponyms associated with the study sites and the German terms *wandern* and *wir* as seeds. *Wandern* and *wir* can be translated as 'walking/hiking' and 'we', respectively, and were selected to find textual descriptions of first person experiences at these locations (i.e. 'As we were hiking along the shores of Lake Zurich. . .'). The BootCaT interface returned around 40 to 70 web pages for each study site, from which we selected the five per site that we judged to be the longest, most landscape-related, first person accounts consisting of full text (as opposed to, for example, only images and their captions). Each web corpus of five texts formed a single 'document' about one of the study locations, and thus our second set of documents, *hiking blogs*, was complete.

In order to create our third set of documents, user-generated content in the form of Flickr tags, a key step was to use a data-driven approach for systematically identifying content associated with the landscape at our study locations (Acheson *et al*. 2017). We chose to use our second set of documents, hiking blogs, as the basis for defining an appropriately-sized region tailored to each study site, which we could then use to query georeferenced Flickr content. Therefore, we first generated a geographical footprint for each study site by using the textual content of each web corpus (or 'document' unit) of hiking blogs, then spatially queried Flickr for images within each resulting footprint. An overview of this process is shown in Figure 2, using the Seealpsee study site as an example. Creating footprints from toponyms contained in documents relating to the study sites enabled us to define the geographic boundaries for querying Flickr content, rather than draw arbitrary boundaries ourselves, or defining a fixed radius around the interview locations that does not take into account the landscape setting at each site.

In a first step, we manually identified all toponyms in each of the 50 texts crawled from the web. This was feasible to do by hand and ensured that we had high quality toponym recognition, as poor performance on this step can propagate downstream (Amitay *et al*. 2004, Purves *et al*. 2007). We then aggregated all toponyms for each study site, resulting in a toponym list for each of the 10 study sites. We queried each of these toponyms for geographic coordinates using the location search feature of the openly available GeoAdmin API (http://api3.geo.admin.ch/), which returns results from the SwissNames3D gazetteer, a nationally produced dataset containing a comprehensive listing of place names in Switzerland, alongside their type and coordinates. To obtain a geographically focused footprint for each study site, we iteratively filtered out candidates that were more than two standard deviations away from the centroid of the points obtained for each toponym (Smith and Crane 2001), stopping when we reached either a footprint dimension threshold or a maximum iteration threshold, both heuristically set. Finally, we calculated the convex hull of the remaining points, resulting in a final set of 10 convex hulls used as footprints to extract georeferenced images. Using the bounding boxes of these convex hulls, we queried the Flickr API for images from these areas. We removed bulk uploads from these initial results by filtering out pictures with the same tags, thereby reducing the bias created by prolific users, often cited as a problem in user-generated content (Hollenstein and Purves 2010). We then further refined the set of pictures by discarding any images outside of our convex hulls. Given the long tail of the tag frequency distribution, with very many idiosyncratic tags only used once, we retained

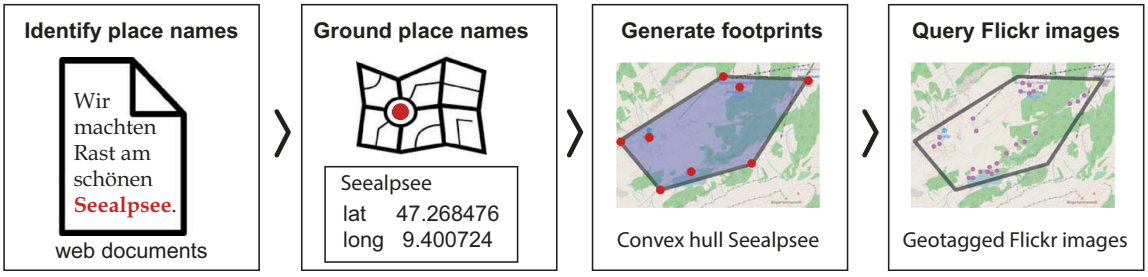


Figure 2. Processing steps to create footprints from web documents and extract georeferenced images.

only tags that were mentioned two times or more per study site for our analysis. Thus, our third data source, Flickr tags, consisted of lists of tags and their frequencies, with one such list or 'document' per study site.

3.3. *Coding scheme*

To annotate the terms contained in our corpora, we devised a coding scheme using an iterative process with open coding informed by a literature review and our own data, followed by structured coding (Crang and Cook 2007). The goal for our coding scheme was that it should reflect different aspects of landscapes contained in our data, ranging from the physical landscape settings to the meanings our respondents ascribe to these settings. From the myriad of theories and conceptualizations of place, place meaning, and sense of place in the literature (Tuan 1977, Low and Altman 1992, Feld and Basso 1996, Twigger-Ross and Uzzell 1996, Jorgensen and Stedman 2001, Williams and Vaske 2003, Cresswell 2006), we selected the concept of place by Agnew (1987), because it includes tangible as well as intangible aspects of people–place relations. The tripartite concept consists of location, locale, and sense of place. The first aspect of location is represented in our data in the form of *toponyms*. The second aspect is locale, or the setting where social life takes place. We further refined the aspect of locale with categories informed from landscape character assessments (Swanwick *et al*. 2002) and thus defined three subcategories: *biophysical landscape elements, cultural landscape elements*, and *perceptual elements*. Biophysical landscape elements contain terms relating to geology, landforms, soil, land cover, flora, fauna, and climate, while cultural landscape elements contain terms referring to land use, settlements, infrastructure, domesticated animals, and anthropogenic objects. Perceptual elements include terms referring to color, touch/feel, sounds, smells, and weather and atmospheric conditions. The third aspect of *sense of place* is represented in our data by mentions of meanings, feelings, memories, as well as terms relating to a sense of attachment, identity, or history of a place or landscape. We thus use sense of place as an umbrella concept, encompassing other concepts such as place identity and place attachment. Additional aspects derived from open coding were *activities* participants associated with a landscape, and *people* in the landscape. Thus, the final coding scheme consisted of seven categories or landscape aspects: toponym, biophysical landscape element, cultural landscape element, perceptual landscape element, sense of place, activity, and people. Based on these seven aspects, the first author applied structured coding to all our data. We then used the coded terms both to compare the data sources in terms of their distribution of aspects, and to select term subsets for cosine similarity comparisons between pairs of documents, described next.

3.4. *Comparing landscape descriptions between data sources*

To quantitatively compare landscape descriptions with respect to the three different data sources and five landscape types, we used cosine similarity to compare documents represented as term vectors (Manning and Schütze 1999). To calculate a cosine similarity between two documents, each document is represented as a vector of N terms, where N is the number of terms appearing in the corpus and each term is weighted by its frequency of occurrence in the respective document. In our study, a free list document consisted of all the terms listed by 30 visitors at a study site, a hiking blog document

consisted of the full text from five web-crawled landscape descriptions relating to a site, and a Flickr tag document consisted of all the tags from images georeferenced at a study site, after the various filtering methods described previously. In a preprocessing step, all documents were cleaned of stop words, and free list entries and hiking blogs were split into their component words. For comparing particular subsets of landscape aspects, we retained only terms that had been coded as pertaining to the specified landscape aspects. The result of this process was a set of 30 term vectors, one per document, forming the basic units for all cosine similarity calculations.

All cosine similarity calculations were performed using the scikit-learn python library (Pedregosa *et al*. 2011). For one set of cosine similarity calculations, where a subset of terms pertaining to particular landscape aspects was used in the term vectors, we assessed the statistical significance of comparisons of groups of cosine similarity values using two-sided Mann-Whitney-U tests at significance level $\alpha = 0.05$.

To illustrate this process, we take as an example the cosine similarities calculated with only biophysical terms, and how we compared data sources based on these cosine similarity values. First, we create a term vector for each document, where each vector has the same length M, where M is the number of words in the entire corpus that were annotated as biophysical landscape aspects. Each entry in the vector for a document represents the frequency of this word in the given document. Next we calculate cosine similarity values between all possible pairs of term vectors, resulting in a 30×30 matrix of values (including comparisons where a document's term vector is compared to itself, resulting in a cosine similarity of 1.0). We then assess whether cosine similarity values were statistically more similar to each other when the pair of documents being compared was from the same data source (e.g. Flickr tags for Oeschinensee vs. Flickr tags for Seealpsee) as opposed to from different data sources (e.g. Flickr tags for Oeschinensee vs. hiking blogs for Seealpsee).

4. Results and interpretation

We first present our findings on the distribution of landscape aspects in different data sources, then our results comparing landscape descriptions across study sites and data sources.

4.1. *Aspects of landscape in different data sources*

After having gathered our three types of landscape descriptions and annotated each term according to the landscape aspect it best represents, we could compare data sources according to their respective content of each landscape aspect. The three data sources differed slightly in the amount of terms they contained. The free lists conducted with 300 visitors contained a mean number of 303 terms per study site (*Mdn* = 307, *SD* = 43.8). In absolute numbers, the 10 web corpora of hiking blogs contained the most terms with a mean of 427 per study site (*Mdn* = 307, *SD* = 97.3). For Flickr image tags, we collected a mean of 332 terms (*Mdn* = 340, *SD* = 156.4). In order to highlight which aspects were more abundant in which data source, we normalized these term counts by the total number of terms contained in each data source.

The distribution of aspects in each data source is presented in Figure 3. Free lists contained the smallest proportion of toponyms of all three data sources, but were a rich source of biophysical properties of landscapes, while also containing high proportions of cultural

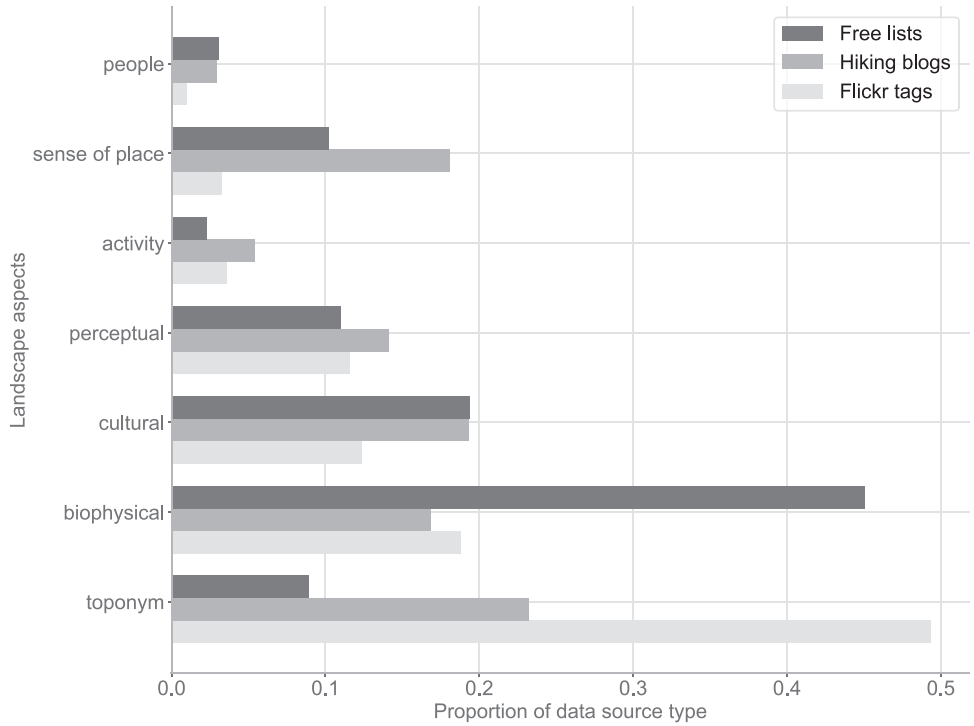


Figure 3. Relative prominence of different landscape aspects depending on the data source.

landscape elements and to some extent, sense of place. The hiking blogs were the richest source of terms relating to sense of place, and contained similar amounts of toponyms, biophysical elements, cultural elements, and perceptual aspects. Flickr tags proved to be a good source of toponyms, accounting for almost half of the annotated content. Flickr tags also contained many terms relating to biophysical and cultural landscape elements, as well as perceptual aspects, but yielded little information on sense of place.

4.2. *Comparing landscape descriptions between data sources*

We found that filtering out toponyms and aspects not related to landscape from all data sources often considerably increased the cosine similarity values between documents from two different data sources. Intuitively, this is consistent with our results from the landscape aspect comparison which show that our three data sources contain very different proportions of toponyms. Thus, removing toponyms and non-landscape aspects is likely to make different data sources more similar, within a study site. For instance, comparing hiking blogs and free lists, before and after removing such aspects, cosine similarity increases for all study sites except River Thur, where it minimally decreases (Table 2).

To statistically assess the influence of the data source on cosine similarity values, we grouped the cosine values of comparisons within the same data source and compared these to cosine values of comparisons between different data sources, repeating this for all our term subsets (e.g. without toponyms, only biophysical landscape aspects, etc.). For example, we compared whether free lists were more similar to other free lists than to image tags, irrespective of the location or landscape type. Our results show that documents of the same data source are significantly more similar than documents from different data sources. Thus, landscape descriptions were more similar within the same data source, irrespective of the

Table 2. Cosine similarity values for the comparison between hiking blogs and free lists.

Study site	Cosine similarity (all terms)	Cosine similarity (landscape terms, no toponyms)	Difference
Oeschinensee	0.2622	0.3339	0.0716
Seealpsee	0.1991	0.2843	0.0852
River Thur	0.2852	0.2848	−0.0003
River Reuss	0.2579	0.2860	0.0281
Robenhuserriet	0.1555	0.1719	0.0164
Ägeriried	0.2220	0.3227	0.1007
Lake Lucerne	0.1343	0.2027	0.0684
Lake Zurich	0.1946	0.2392	0.0446
Lägern	0.1713	0.2156	0.0443
Pfannenstiel	0.1661	0.2165	0.0504

study site or landscape type they were located in. For all term subsets, we obtained statistically significant differences (two-sided Mann-Whitney-U), except for the biophysical landscape terms for the hiking blog comparison (Table 3).

4.3. *Comparisons between landscape types*

Our second comparison was between landscape types, that is, comparing documents from one landscape type (in different data sources) against documents from all other landscape types. For example, we tested whether mountain landscapes in Flickr tags, hiking blogs, and free lists were more similar to each other than they were to all descriptions of other landscape types. Here, we found that most comparisons were not significant (Table 4). Given the results from the previous section, this result was highly probable, since documents of a particular data source were more similar to documents of the same data source than to documents from a different data source, irrespective of the landscape type. However, there were a few significant comparisons, notably for river landscapes, specifically for all terms coded as landscape aspects (both with and without toponyms) and only terms coded as biophysical aspects (Table 4). The high cosine similarity values between river landscapes seems to be driven by the similarity of terms used for biophysical properties (e.g. water, trees, river), leading us to believe that the limited vista space for river landscapes limits the inclusion of terms from other semantic fields, which could have been a potential driver for this similarity.

Table 3. Mann-Whitney-U values for within-document-type vs. across-document-type cosine similarity comparisons (statistically significant results in bold).

	All terms (including non-landscape)	All landscape terms	Landscape terms, no toponyms	Biophysical terms only	Sense of place terms only
Free lists	U = 8970, **p< .001**	U = 8873, **p< .001**	U = 8480, **p< .001**	U = 7429, **p< .001**	U = 8078, **p< .001**
Hiking blogs	U = 8898, **p< .001**	U = 5568, **p = .013**	U = 5921, **p< .001**	U = 4273, p = .598	U = 5834, **p = .002**
Flickr tags	U = 9000, **p< .001**	U = 8998, **p< .001**	U = 8374, **p< .001**	U = 7926, **p< .001**	U = 7190, **p< .001**

Table 4. Mann-Whitney-U values for within-landscape-type vs. across-landscape-type cosine similarity comparisons (statistically significant results in bold).

	All terms (including non-landscape)	All landscape terms	Landscape terms, no toponyms	Biophysical terms only	Sense of place terms only
Mountain landscape	U = 1344, p = .120	U = 1402, p = .058	U = 1344, p = .106	U = 1337, p = .131	U = 1444, **p = .031**
River landscape	U = 1405, p = .056	U = 1512, **p = .011**	U = 1680, **p < .001**	U = 1936, **p < .001**	U = 1200, p = .479
Moor landscape	U = 1161, p = .635	U = 1189, p = .523	U = 1126, p = .789	U = 1096, p = .927	U = 993, p = .606
Lake landscape	U = 1200, p = .481	U = 1267, p = .272	U = 1375, p = .083	U = 1676, **p < .001**	U = 913.5, p = .300
Hill landscape	U = 1206, p = .460	U = 1263, p = .282	U = 1150, p = .682	U = 1228, p = .385	U = 908, p = .302

4.4. *Comparisons between landscape types within a data source*

As a final step, to control for the influence of the data source, we compared landscape types within a data source. For example, we took cosine values of comparisons between pairs of documents of Flickr tags from the same landscape type (e.g. hill landscape vs. hill landscape) and compared them to pairs of documents of Flickr tags from different landscapes types (e.g. mountain vs. river, hill vs. moor, lake vs. moor, and so on). We found that, within our three data sources, documents from the same landscape type were significantly more similar than documents from different landscape types for two-term subsets: landscape aspects excluding toponyms, and only biophysical aspects (Table 5). However, with the sense of place term subset, there were no significant differences between descriptions within and between landscape types.

Table 5. Mann-Whitney-U values for within-landscape-type vs. across-landscape-type cosine similarity comparisons, within a data source (statistically significant results in bold).

	All terms (including non-landscape)	All landscape terms	Landscape terms, no toponyms	Biophysical terms only	Sense of place terms only
Free lists: same vs. different landscape types	U = 197, **p < .001**	U = 197, **p < .001**	U = 199, **p < .001**	U = 194, **p < .001**	U = 118, p = .527
Hiking blogs: same vs. different landscape types	U = 117, p = .551	U = 163, **p = .024**	U = 160, **p = .032**	U = 167, **p = .016**	U = 140, p = .154
Flickr tags: same vs. different landscape types	U = 131, p = .271	U = 128, p = .321	U = 156, **p = .045**	U = 181, **p = .004**	U = 93, p = .814

5. Discussion

In this study, we devised and applied a methodology that is based on the combination of both empirical elicitation of information about landscapes with participants in a landscape setting and more data-driven extraction approaches informed from text retrieval. The discussion is structured around the three main contributions of this article: our integrated methodology for gathering landscape descriptions, our finding that we can distinguish between formally recognized landscape types based on landscape descriptions, and our application of GIScience methods to the domain of landscape characterization.

5.1. *An integrated methodology for gathering landscape descriptions from the public*

Conducting free listing in outdoor settings, we elicited landscape categories and other associations with landscape from the public, including notions of scenicness, relaxation, and identity. This method from cognitive psychology has proven to be transferable to the domain of landscape and was empirically tested (Williams *et al*. 2012, Wartmann 2015). Our study confirms the usefulness of free listing with participants for eliciting landscape categories and associated terms, including information pertaining to sense of place. We then crawled the web for landscape descriptions, based on seeds that included the place name of the study site, to create a small corpus of full text descriptions. Though different seeds may be tested in further research, our choice of seeds provided sufficient content for each study site, enabling us to make a selection based on the suitability of the text for our use case. We purposefully focused on accounts of people having visited the landscape and writing about their first-hand experience, excluding texts such as descriptions on tourism websites. As we used these full text descriptions to create geographically focused footprints to query Flickr photos, we restricted ourselves to a small number of texts so that we could manually annotate all toponyms, thus achieving a gold standard in toponym recognition. This time-consuming task could be avoided in the future by implementing automated toponym recognition and resolution (Amitay *et al*. 2004). However, these are challenging tasks made harder still with semi-formal texts in German like our hiking blogs, and authors who liberally use vernacular and idiosyncratic spellings of place names (Augenstein *et al*. 2017), common in Switzerland with its multitude of oral German dialects. By basing our queries on site-specific footprints, we gathered images taken in focused geographical areas, rooted in the way a landscape was described in text, thus going further than methods such as queries based on a simple distance criterion.

We thus created three data sources (free lists, hiking blogs, and Flickr tags) as the basis for further comparisons. Overall, our approach proved well suited to study highly frequented landscapes. However, it would be more challenging to apply to landscapes with fewer visitors, where the time and costs to conduct interviews and free listings increase, and the availability of web content (blogs, social media) decreases, given the highly unequal distribution of user-generated content across space (Antoniou *et al*. 2010). If our intention is to eventually create data layers containing descriptions with full spatial coverage, this is a severe limitation of passively crowdsourced information. Thus, alternative approaches are to include actively crowdsourced information through citizen science initiatives (Connors *et al*. 2012, Haklay 2013). For instance, platforms where users upload full-text landscape descriptions for public use, such as in the Geograph Britain and Ireland project (http://www.geograph.org.uk/), can provide semantically rich crowdsourced information focused on landscape.

5.2. *Comparing landscape descriptions between settings and data sources*

Our results show that landscape descriptions from the same data source were more similar than between data sources, also for different landscape types. This indicates that there are considerable differences in the data we capture through these approaches in terms of the lexical aspects, resulting in low cosine similarities for comparisons between data sources (e.g. Flickr tags with hiking blogs). However, the results from coding

aspects of landscape confirm that these differences are not merely lexical, but also semantic. For instance, while Flickr tags contained high percentages of toponyms but less content on sense of place, hiking blogs contained more sense of place content. The relatively high proportion of content in hiking blogs related to sense of place compared with Flickr tags and free lists could be expected, because narratives used for hiking blogs may lend themselves better to expressing sense of place than single words used as tags or in free lists. However, free-listing experiments also result in some content relating to sense of place, despite the elicitation question not aimed at documenting such content (Wartmann and Purves, in press). The fact that people list terms related to sense of place in free-listing tasks at all may be linked to memory retrieval processes, where people first list landscape features, and then also list feelings and meanings associated with the landscape (Wartmann 2015). Thus, we argue that rather than being interchangeable, these data sources are complementary, each highlighting different aspects of landscape. Indeed, this combination of approaches is particularly promising for applications that aim to include multiple perspectives on landscapes, including expert and nonexpert opinions on landscape characterization (Dalglish and Leslie 2016). The limitations of our approaches are that each of them reaches a particular subset of the population (e.g. people visiting a place on a sunny summer day and being interviewed, others visiting the same place at a different time and writing about it on the web or uploading pictures), and that the content we documented is perhaps reflective of that part of the population, but not of others. For comparing descriptions between landscape types, we found that zooming into a single data source (e.g. hiking blogs), two sites in the same landscape type were lexically more similar if we excluded toponyms from our data. By narrowing the term subset further down to biophysical aspects of landscape only, we observed this effect strongly across all data sources. This finding supports our hypothesis that people, at least within a cultural-linguistic group, describe landscapes differently in terms of their biophysical properties, such as whether one is a mountain landscape characterized by rocks, cliffs, and crevasses, or a river landscape with a river flowing through woods and farmlands. Interestingly, we found no differences between landscape types in the terms relating to sense of place. This finding suggests people's description of their experience varies to a lesser extent between different landscape types. Landscapes that are generally perceived as natural and are visited for recreational purposes may thus evoke similar feelings of identity, relaxation, and a connection to nature (in other words cultural ES), irrespective of the biophysical composition of the landscape. This is in accordance with research in environmental psychology finding pronounced differences between recreation provided by urban and natural landscapes (Velarde *et al*. 2007), but fewer differences in recreational effects between different natural landscapes such as forests, hills, and moors, with the exception of coastal landscapes that provided more restoration (White *et al*. 2013). The methodology we describe in this article thus has a range of potential applications, which we outline next.

5.3. *Relevance and potential applications*

This study highlights the potential of adapting methods from GIScience and geographic information retrieval (GIR) to the domain of landscape characterization, with possible applications of refining landscape character assessments and landscape typologies using

data gathered from the public. Indeed, by combining data-driven approaches from GIR with *in situ* elicitation of landscape character, we are contributing toward enriching representations of geographic information. Our methods centered around collecting and comparing text data that contained terms for landscape features and semantically related content. We argue that integrating such information grounded in individual perceptions and language with spatial data potentially allows us to include more situated knowledge and subjective perceptions into existing, typically expert-driven, data. Our work is in line with previous efforts for bringing to the fore meanings of geographic information that are relevant to many applications of GIS (Gahegan and Pike 2006). Furthermore, we showed how different data sources (free lists, hiking blogs, and Flickr tags) differ in their landscape-related content. Combining these different sources thus provides a more holistic data basis (e.g. for landscape character assessments) than any of these data sources on its own.

With the arrival of new forms of crowdsourced data in ever increasing volumes, this article focused on the question of how we can 'dig into this data avalanche' (Miller 2010) to answer questions relevant to landscape research. So far, most of the research on novel data about places and landscapes has focused on the enumeration of the quantity of user-generated content relating to certain areas (Nahuelhual *et al*. 2013, Tenerelli *et al*. 2016). If semantics were taken into account at all, the focus was on retrieving content that related to a fixed set of predefined keywords (van Zanten *et al*. 2016), or in some cases, applying automated methods to define topics emerging from large volumes of data (Jenkins *et al*. 2016). In this study, we focused on the content, using the link between text and geographical space to access and harness different data sources. Our work builds on previous research revolving around the extraction of place semantics in user-generated content (Rattenbury and Naaman 2009, Hollenstein and Purves 2010, Capineri 2016). Instead of characterizing a single place through different data sources (Capineri 2016), we took a comparative approach, working out the similarities and differences in descriptions between different landscapes in different data sources. We did this both in a qualitative way by looking more closely at the semantics, as well as lexically through quantitative text comparisons. We showed that, by carefully sifting the 'avalanche' and creating focused corpora, we can apply semi-automated methods for processing this information that lead to more than just a shallow reading, yet which are scalable and yield sufficient landscape-related content to be interesting for practical applications. Such practical applications include, for example, the assessment of indicators for cultural landscape values (Bieling *et al*. 2014).

Our work in this field is novel in that it links the empirical method of free listing with digital traces in the form of web-crawled documents and georeferenced images. While methodological challenges remain in adapting methods to larger scales (and potentially larger volumes of data), we demonstrate the potential of combining methodologies to overcome disciplinary boundaries and provide multiple perspectives on the complex phenomenon of the relation between people and landscapes.

6. Conclusions

By using both traditional empirical methods as well as emerging crowdsourced data, our approach puts people and their perception, experience, and appreciation of landscapes at the center. Using three different data sources, we highlight the potential of gathering

landscape descriptions from the bottom-up, and show how these sources allow us to distinguish landscape types based on the description of the biophysical landscape features. Such an approach constitutes a step toward including the views of a wider public, integrating multiple perspectives, and adding semantics to the typically expert-based spatial data collected for landscape assessments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

We thank the cogito foundation for the financial support through the project 'How language shapes our sense of place', grant no. 15-129-R.

ORCID

Flurina M. Wartmann http://orcid.org/0000-0003-4788-2963

References

Acheson, E., Wartmann, F.M., and Purves, R.S., 2017. Generating spatial footprints from hiking blogs. *In*: P. Fogliaroni, A. Ballatore, and E. Clementini, eds. *Proceedings of workshops and posters at the 13th International Conference on Spatial Information Theory (COSIT 2017). Lecture Notes in Geoinformation and Cartography*. Cham: Springer, 5–7.

Agnew, J.A., 1987. *Place and politics. The geographical mediation of state and society*. Boston: Allen & Unwin.

Amitay, E., *et al*., 2004. Web-a-where: geotagging web content. *In*: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* ACM, 273–280.

Antoniou, V., Morley, J., and Haklay, M., 2010. Web 2.0 geotagged photos: assessing the spatial dimension of the phenomenon. *Geomatica*, 64(1), 99–110.

ARE, 2011a. *Die Landschaften der Schweiz. Landschaftstypologie Schweiz*. Technical report, ARE, Berne, Switzerland.

ARE, 2011b. *Landschaftstypologie Schweiz Teil 1, Ziele, Methode und Anwendung*. Technical report, ARE, Berne, Switzerland.

Augenstein, I., Derczynski, L., and Bontcheva, K., 2017. Generalisation in named entity recognition: a quantitative analysis. *Computer Speech & Language*, 44, 61–83. doi:10.1016/j.csl.2017.01.012

Baroni, M. and Bernardini, S., 2004. BootCaT: bootstrapping corpora and terms from the web. *In*: *LREC - International Conference on Language Resources and Evaluation* ELRA.

Bieling, C., *et al*., 2014. Linkages between landscapes and human well-being: an empirical exploration with short interviews. *Ecological Economics*, 105, 19–30. doi:10.1016/j.ecolecon.2014.05.013

Brabyn, L., 1996. Landscape classification using GIS and national digital databases. *Landscape Research*, 21(3), 277–300. doi:10.1080/01426399608706493

Bunce, R., *et al*. 1996. Land classification for strategic ecological survey. *Journal of Environmental Management*, 47(1), 37–60. doi:10.1006/jema.1996.0034

Butler, A., 2016. Dynamics of integrating landscape values in landscape character assessment: the hidden dominance of the objective outsider. *Landscape Research*, 41(2), 239–252. doi:10.1080/01426397.2015.1135315

Capineri, C., 2016. Kilburn high road revisited. *Urban Planning*, 1(2), 128. doi:10.17645/up.v1i2.614

Caspersen, O.H., 2009. Public participation in strengthening cultural heritage: the role of landscape character assessment in Denmark. *Geografisk Tidsskrift-Danish Journal of Geography*, 109(1), 33–45. doi:10.1080/00167223.2009.10649594

Comber, A., 2013. Comparing expert and non-expert conceptualisations of the land: an analysis of crowdsourced land cover data. *In*: T. Tenbrink, *et al*., eds. *International Conference on Spatial Information Theory. COSIT 2013: Spatial Information Theory. Lecture Notes in Computer Science*. Vol. *8116*. Cham: Springer, 243–260.

Connors, J.P., Lei, S., and Kelly, M., 2012. Citizen science in the age of neogeography: utilizing volunteered geographic information for environmental monitoring. *Annals of the Association of American Geographers*, 102(6), 1267–1289. doi:10.1080/00045608.2011.627058

Costanza, R., *et al*. 1997. The value of the world's ecosystem services and natural capital. *Nature*, 387(6630), 253–260. doi:10.1038/387253a0

Council of Europe, 2000. *European Landscape Convention*. Technical report, Florence.

Crang, M. and Cook, I., 2007. *Doing ethnographies*. London: Sage.

Cresswell, T., 2006. *Place*. Malden, MA: Blackwell Pub.

Dalglish, C. and Leslie, A., 2016. A question of what matters: landscape characterisation as a process of situated, problem-orientated public discourse. *Landscape Research*, 41(2), 212–226. doi:10.1080/01426397.2015.1135319

Daniel, T.C. and Vining, J. 1983. Methodological issues in the assessment of landscape quality. *In*: I. Altman and J. Wohlwill, eds. *Behavior and the natural environment. Human behavior and environment (Advances in theory and research)*. Boston, MA: Springer, Vol. 6, 39–84.

de Groot, R., *et al*. 2010. Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. *Ecological Complexity*, 7(3), 260–272. doi:10.1016/j.ecocom.2009.10.006

Derungs, C. and Purves, R.S., 2013. From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, 28(6), 1272–1293. doi:10.1080/13658816.2013.772184

Derungs, C. and Purves, R.S., 2016. Characterising landscape variation through spatial folksonomies. *Applied Geography*, 75, 60–70. doi:10.1016/j.apgeog.2016.08.005

Dong, J., *et al*. 2003. Remote sensing estimates of boreal and temperate forest woody biomass: carbon pools, sources, and sinks. *Remote Sensing of Environment*, 84(3), 393–410. doi:10.1016/S0034-4257(02)00130-X

Dunkel, A., 2015. Visualizing the perceived environment using crowdsourced photo geodata. *Landscape and Urban Planning*, 142, 173–186. doi:10.1016/j.landurbplan.2015.02.022

Edwardes, A. and Purves, R., 2007. A theoretical grounding for semantic descriptions of place. *In*: J. M. Ware and G.E. Taylor, eds. *Web and Wireless Geographical Information Systems. 7th International Symposium, W2GIS 2007, Cardiff, UK, November 28-29, 2007. Proceedings*. Berlin Heidelberg: Springer, 106–120.

Feld, S. and Basso, K., eds., 1996. *Senses of place*. Santa Fe, New Mexico: School of American Research Press.

Figueroa-Alfaro, R.W. and Tang, Z., 2017. Evaluating the aesthetic value of cultural ecosystem services by mapping geo-tagged photographs from social media data on Panoramio and Flickr. *Journal of Environmental Planning and Management*, 60(2), 266–281. doi:10.1080/09640568.2016.1151772

Gahegan, M. and Pike, W., 2006. A situated knowledge representation of geographical information. *Transactions in GIS*, 10(5), 727–749. doi:10.1111/tgis.2006.10.issue-5

Gerçek, D., Toprak, V., and Strobl, J., 2011. Object-based classification of landforms based on their local geometry and geomorphometric context. *International Journal of Geographical Information Science*, 25(6), 1011–1023. doi:10.1080/13658816.2011.558845

Giannakopoulou, L., *et al*. 2013. From compasses and maps to mountains and territories: experimental results on geographic cognitive categorization. *In*: M.M. Raubal, D.M. Mark, and A.U. Frank, eds. *Cognitive and Linguistic Aspects of Geographic Space. Lecture Notes in Geoinformation and Cartography*. Berlin: Springer, 63–81.

Gregory, I.N. and Hardie, A., 2011. Visual GISting: bringing together corpus linguistics and geographical information systems. *Literary and Linguistic Computing*, 26(3), 297–314. doi:10.1093/llc/fqr022

Grothe, C. and Schaab, J., 2009. Automated footprint generation from geotags with Kernel density estimation and support vector machines. *Spatial Cognition & Computation*, 9(3), 195–211. doi:10.1080/13875860903118307

Gschwend, C. and Purves, R.S., 2012. Exploring geomorphometry through user generated content: comparing an unsupervised geomorphometric classification with terms attached to georeferenced images in Great Britain. *Transactions in GIS*, 16(4), 499–522. doi:10.1111/j.1467-9671.2012.01307.x

Guerrero, P., *et al*. 2016. Revealing cultural ecosystem services through Instagram images: the potential of social media volunteered geographic information for urban green infrastructure planning and governance. *Urban Planning*, 1(2), 1. doi:10.17645/up.v1i2.609

Haklay, M., 2013. Citizen science and volunteered geographic information: overview and typology of participation. *In*: M.H. Palmer and S. Kraushaar, eds. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Dordrecht: Springer Netherlands, 105–122.

Hollenstein, L. and Purves, R., 2010. Exploring place through user-generated content: using Flickr to describe city cores. *Journal of Spatial Information Science*, 1, 21–48. doi:10.5311/JOSIS.2010.1.3

Iwahashi, J. and Pike, R.J., 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*, 86(3–4), 409–440. doi:10.1016/j.geomorph.2006.09.012

Jenkins, A., *et al*. 2016. Crowdsourcing a collective sense of place. *Plos One*, 11(4), e0152932. doi:10.1371/journal.pone.0152932

Jessel, B., 2006. Elements, characteristics and character – information functions of landscapes in terms of indicators. *Ecological Indicators*, 6(1), 153–167. doi:10.1016/j.ecolind.2005.08.009

Johnson, L. and Hunn, E., 2010. *Landscape ethnoecology*. New York: Berghahn Books.

Jorgensen, B.S. and Stedman, R.C., 2001. Sense of place as an attitude: lakeshore owners attitudes toward their properties. *Journal of Environmental Psychology*, 21(3), 233–248. doi:10.1006/jevp.2001.0226

Kienast, F., *et al*., 2015. The Swiss landscape monitoring program – A comprehensive indicator set to measure landscape change. *Ecological Modelling*, 295, 136–150. doi:10.1016/j.ecolmodel.2014.08.008

Kirchhoff, T., 2012. Pivotal cultural values of nature cannot be integrated into the ecosystem services framework. *Proceedings of the National Academy of Sciences*, 109(46), E3146–E3146. doi:10.1073/pnas.1212409109

Low, S.M. and Altman, I., eds., 1992. *Place attachment*. New York, N.Y: Plenum Press.

MA, 2005. *Millennium ecosystem assessment*. Washington D.C: World Resources Institute.

Manning, C.D. and Schütze, H., 1999. *Foundations of statistical natural language processing*. Vol. 999, Cambridge: MIT Press.

Mark, D.M., *et al*. 2011. *Landscape in language*. Amsterdam/Philadelphia: John Benjamins Publishing.

Mark, D.M., Smith, B., and Tversky, B. 1999. Ontology and geographic objects: an empirical study of cognitive categorization. *In*: C. Freksa, ed. *International Conference on Spatial Information Theory. COSIT 1999: Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science. Lecture Notes in Computer Science*. Berlin: Springer, Vol. *1661*, 283–298.

Mark, D.M. and Turk, A.G. 2003. Landscape categories in Yindjibarndi: ontology, environment and language. *In*: W. Kuhn, M. Worboys, and S. Timpf, eds. *Spatial Information Theory. Foundations of Geographic Information Science. COSIT 2003. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, Vol. *2825*, 28–45.

Miller, H.J., 2010. The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, 50(1), 181–201. doi:10.1111/jors.2010.50.issue-1

Mücher, C.A., *et al*. 2010. A new European Landscape Classification (LANMAP): a transparent, flexible and user-oriented methodology to distinguish landscapes. *Ecological Indicators*, 10(1), 87–103. doi:10.1016/j.ecolind.2009.03.018

Nahuelhual, L., *et al*., 2013. Mapping recreation and ecotourism as a cultural ecosystem service: an application at the local level in Southern Chile. *Applied Geography*, 40, 71–82. doi:10.1016/j.apgeog.2012.12.004

Natural England, 2014. *Landscape and seascape character assessments*. Technical report.
Niesterowicz, J., Stepinski, T., and Jasiewicz, J., 2016. Unsupervised regionalization of the United States into landscape pattern types. *International Journal of Geographical Information Science*, 30 (7), 1450–1468. doi:10.1080/13658816.2015.1134796
Oteros-Rozas, E., *et al*., 2017. Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites. *Ecological Indicators*. doi:10.1016/j.ecolind.2017.02.009
Pedregosa, F., *et al*. 2011. Scikit-learn: {Machine} {Learning} in {Python}. *Journal of Machine Learning Research*, 12, 2825–2830.
Pickles, J., 1995. *Ground truth*. New York: Guilford Press.
Pitt, D., 1989. The attractiveness and use of aquatic environments as outdoor recreation places. *In*: I. Altman and E.H. Zube, eds. *Public Spaces and Places*. New York, NY: Plenum Press, 217–254.
Purves, R.S., *et al*. 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7), 717–745. doi:10.1080/13658810601169840
Rattenbury, T. and Naaman, M., 2009. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3(1), 1–30. doi:10.1145/1462148
Richards, D.R. and Friess, D.A., 2015. A rapid indicator of cultural ecosystem service usage at a fine spatial scale: content analysis of social media photographs. *Ecological Indicators*, 53, 187–195. doi:10.1016/j.ecolind.2015.01.034
Rorissa, A., 2008. User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory. *Information Processing & Management*, 44(5), 1741–1753. doi:10.1016/j.ipm.2008.03.004
Sarlöv Herlin, I., 2016. Exploring the national contexts and cultural ideas that preceded the landscape character assessment method in England. *Landscape Research*, 41(2), 175–185. doi:10.1080/01426397.2015.1135317
Scott, A., 2002. Assessing public perception of landscape: the LANDMAP experience. *Landscape Research*, 27(3), 271–295. doi:10.1080/01426390220149520
Scott, A., 2003. Assessing public perception of landscape: from practice to policy. *Journal of Environmental Policy & Planning*, 5(2), 123–144. doi:10.1080/1523908032000121193
Smith, B. and Mark, D.M., 2003. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B: Planning and Design*, 30, 411–427. doi:10.1068/b12821
Smith, D.A. and Crane, G., 2001. Disambiguating geographic names in a historical digital library. *In*: P. Constantopoulos and I.T. Sølvberg, eds. *International Conference on Theory and Practice of Digital Libraries. Lecture Notes in Computer Science 2163*. Heidelberg: Springer, 127–136.
Swanwick, C., 2002. *Landscape Character Assessment Guidance for England and Scotland*. Technical report, countryside agency and Scottish natural heritage.
Swanwick, C., 2004. The assessment of countryside and landscape character in England: an overview. *In*: K. Bishop and A. Phillips, eds. *Countryside planning: new approaches to management and conservation*. Camden, London: Earthscan, 109–124.
Swanwick, C., Bingham, L., and Parfitt, A., 2002. Topic paper 3: landscape character assessment. How stakeholders can help. *Landscape Character Assessment Guidance for England and Scotland. The Countryside Agency and Scottish Natural Heritage*. Available from: http://publications.naturalengland.org.uk/publication/4902400292814848.
Tenerelli, P., Demšar, U., and Luque, S., 2016. Crowdsourcing indicators for cultural ecosystem services: A geographically weighted approach for mountain landscapes. *Ecological Indicators*, 64, 237–248. doi:10.1016/j.ecolind.2015.12.042
Tuan, Y.F., 1977. *Space and place*. Minneapolis: University of Minnesota press.
Tversky, B. and Hemenway, K., 1983. Categories of environmental scenes. *Cognitive Psychology*, 15 (1), 121–149. doi:10.1016/0010-0285(83)90006-3
Twigger-Ross, C.L. and Uzzell, D.L., 1996. Place and identity processes. *Journal of Environmental Psychology*, 16(3), 205–220. doi:10.1006/jevp.1996.0017

Van Eetvelde, V. and Antrop, M., 2009a. A stepwise multi-scaled landscape typology and characterisation for trans-regional integration, applied on the federal state of Belgium. *Landscape and Urban Planning*, 91(3), 160–170. doi:10.1016/j.landurbplan.2008.12.008

Van Eetvelde, V. and Antrop, M., 2009b. Indicators for assessing changing landscape character of cultural landscapes in Flanders (Belgium). *Land Use Policy*, 26(4), 901–910. doi:10.1016/j.landusepol.2008.11.001

van Zanten, B.T., *et al*. 2016. Continental-scale quantification of landscape values using social media data. *Proceedings of the National Academy of Sciences*, 113(46), 12974–12979. doi:10.1073/pnas.1614158113

Velarde, M., Fry, G., and Tveit, M., 2007. Health effects of viewing landscapes – landscape types in environmental psychology. *Urban Forestry & Urban Greening*, 6(4), 199–212. doi:10.1016/j.ufug.2007.07.001

Wang, W. and Stewart, K., 2015. Spatiotemporal and semantic information extraction from web news reports about natural hazards. *Computers, Environment and Urban Systems*, 50, 30–40. doi:10.1016/j.compenvurbsys.2014.11.001

Warnock, S. and Griffiths, G., 2015. Landscape characterisation: the living landscapes approach in the UK. *Landscape Research*, 40(3), 261–278. doi:10.1080/01426397.2013.870541

Wartmann, F. and Purves, R., 2017. What's (not) on the map: landscape features from participatory sketch mapping differ from local categories used in Language. *Land*, 6(4), 79. doi:10.3390/land6040079

Wartmann, F.M., 2015. More than a list: what outdoor free listings of landscape categories reveal about commonsense Geographic concepts and memory search strategies. *In*: S.I. Fabrikant, *et al*., eds. *Spatial Information Theory. Lecture Notes in Computer Science*. Vol. 9368. Heidelberg: Springer, 224–243.

Wartmann, F. M., & Purves, R. S. (in press). Investigating sense of place as a cultural ecosystem service in different landscapes through the lens of language. *Landscape and Urban Planning*.

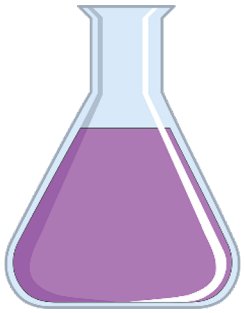
White, M.P., *et al*., 2013. Feelings of restoration from recent nature visits. *Journal of Environmental Psychology*, 35, 40–51. doi:10.1016/j.jenvp.2013.04.002

Williams, D.R. and Vaske, J.J., 2003. The measurement of place attachment: validity and generalizability of a psychometric approach. *Forest Science*, 49(6), 830–840.

Williams, M., Kuhn, W., and Painho, M., 2012. The influence of landscape variation on landform categorization. *Journal of Spatial Information Science*. doi:10.5311/JOSIS.2012.5.107

Yoshimura, N. and Hiura, T., 2017. Demand and supply of cultural ecosystem services: use of geotagged photos to map the aesthetic value of landscapes in Hokkaido. *Ecosystem Services*, 24, 68–78. doi:10.1016/j.ecoser.2017.02.009

Zube, E.H., 1984. Themes in Landscape Assessment Theory. *Landscape Journal*, 3(2), 104–110. doi:10.3368/lj.3.2.104



Paper IV

Extracting and modeling geographic information from scientific articles

Summary

This paper describes a text-to-space pipeline built to extract relevant locations from two corpora of scientific articles: a biomedical and an ecological corpus. The pipeline uses freely available tools, is minimally customized for the corpus domain, and gives promising results which can be useful in the context of a meta-analysis or a geographically-aware search/filtering task.

Contribution of the PhD candidate

I designed, developed, and tested the text-to-space pipeline, wrote all the analysis code in Python, drafted the manuscript, and prepared all the maps and figures.

Citation

Submitted to PLOS ONE: Acheson, E., and Purves, R. S. (2019). Extracting and modeling geographic information from scientific articles.

Extracting and modeling geographic information from scientific articles

Elise Acheson[1*], Ross S. Purves[1]

1 Department of Geography, University of Zurich, Zurich, Switzerland

* elise.acheson@geo.uzh.ch

Abstract

Scientific articles often contain relevant geographic information such as where field work was performed or where patients were treated. Most often, this information appears in the full-text article contents as a description in natural language including place names, with no accompanying machine-readable geographic metadata. Automatically extracting this geographic information could help conduct meta-analyses, find geographical research gaps, and retrieve articles using spatial search criteria. Research on this problem is still in its infancy, with many works manually processing corpora for locations and few cross-domain studies. In this paper, we develop a fully automatic pipeline to extract and represent relevant locations from scientific articles, applying it to two varied corpora. We obtain good performance, with full pipeline precision of 0.84 for an environmental corpus, and 0.78 for a biomedical corpus. Our results can be visualized as simple global maps, allowing human annotators to both explore corpus patterns in space and triage results for downstream analysis. Future work should not only focus on improving individual pipeline components, but also be informed by user needs derived from the potential spatial analysis and exploration of such corpora.

Introduction

Geographical information permeates the written world, appearing as place names or place descriptions in texts including news articles, blog posts, social media content, historical documents, and scientific articles. Research on extracting geographical information from text has often focused on news articles [1–3] and social media content [4–6], with surprisingly limited attention being directed towards the increasing number of published scientific articles. Indeed, with each passing year, scientists face an ever-growing stack of scientific articles to sort through, read, understand, and build upon. Many of these articles contain important geographical information: perhaps soil samples were taken from a certain region, patients were treated in a particular hospital, or interviews were conducted in a village or neighborhood. Currently, researchers must manually sift through article contents to identify any relevant locations, a time-consuming process. Furthermore, linking these textual place descriptions to spatial representations (such as point coordinates, a bounding box, or a polygonal region) requires significant additional work and should ideally respect the scale and precision of locations described in the text. Despite discussions about the need to develop and adopt metadata reporting standards for geographic information [7–9], the vast majority of scientific articles continue to be published without any accompanying machine-readable spatial data, though geographic information often appears in the article contents in

textual form. The ability to automatically extract and spatially represent this geographic information would enable researchers to organize and find information using not just keywords but also spatial criteria, as is done for other types of text using Geographic Information Retrieval (GIR) techniques [10]. Organizing and visualizing scientific corpora by space would facilitate geographically-aware meta-analyses [11], enable studies to be cross-referenced by location [12], and allow for the discovery of geographical research gaps such as understudied regions in a particular scientific discipline [13, 14].

Though scientific articles have become a frequent object of study for researchers, common research objectives are to analyze and visualize (often large) article collections [15–17], and to extract or summarize specific information from publications through text mining, usually in a particular domain such as biomedical research [18, 19]. On the one hand, many scientific corpus analyses consider geography, but focus on author locations which are easier to extract from articles [16, 20], and on the other hand, many specialized text mining tools go beyond article metadata and into full-text processing, but don't give special treatment to geographical information. Meanwhile, extracting and representing meaningful geographical locations such as study sites from scientific articles remains a challenging and understudied problem. Most published works on this problem identify relevant locations from text manually [12, 14, 21, 22], and only few works tackle the problem using a scalable, automatic approach [9, 23, 24]. When automatic approaches are used, they are constrained in their applicability, either by only extracting geographic coordinates [25, 26], by not utilizing the full-text of articles [13], or by performing overly poorly on the full-text [9]. Furthermore, the corpora used remain limited both in size and disciplinary focus, potentially limiting the wider applicability of the techniques and findings.

A long-standing related and relevant stream of work that has recently been applied to scientific articles is the detection and disambiguation of place names (toponyms), a task known as toponym recognition and resolution. One recent strand of work has concentrated around an annotated corpus related to phylogeography [27]. This work includes a series of publications [27–30] and a SemEval-2019 task called 'Toponym Resolution in Scientific Papers'[1] [31]. However, these research efforts focus on identifying *all* toponym mentions within the main text of an article, rather than a subset of relevant locations representing, for example, where a study was conducted. This means that annotated toponyms in this phylogeography corpus include toponyms listed alongside company locations (for chemicals or products used in a study) as well as toponyms mentioned in the context of scientific background. The present work focuses on a different, albeit related, task: automatically extracting and geographically representing meaningful or relevant locations from scientific articles such as study sites, patient treatment locations, and sample locations. These are almost always a (relatively small) relevant subset of the textual locations or toponyms that appear in the article contents, and thus our task relates more closely to finding the geographic scope of text documents [32–34] than to performing comprehensive toponym resolution on each document [35]. Our goal in this paper is rather to replicate what a human annotator would extract from a scientific article for the purposes of a meta-analysis, or what an author would potentially include as geographical metadata for a submitted article.

Indeed, an important part of processing scientific articles is not only to detect locations, but also to ignore irrelevant locations such as locations in references, locations indicating where a company providing commercial products is based, or locations appearing in expressions such as 'the Declaration of Helsinki'. The presence of irrelevant place names throughout scientific articles is cited as a major obstacle to automatically extracting study sites using place names in [26] and affected performance and processing

[1] https://competitions.codalab.org/competitions/19948

decisions in [13,24]. In our task, each individual place name or toponym mention (which we refer to as location mentions since named locations like hospitals and universities are of interest to us but not necessarily considered to be 'toponyms') appearing in text is not equally important, including repeated locations, as long as the correct study locations are captured, as measured through precision and recall.

In this paper, we develop a fully automatic pipeline which starts from a collection of scientific articles and their PDFs and outputs a set of location strings and their sentence context, as well as structured information and a geometric representation for each string (Fig 1). We use two contrasting corpora from two different research domains: 1. a highly spatial ecological research corpus of articles relating to orchards, with most including study site descriptions, and some including maps and coordinates, and 2. a less spatial biomedical corpus of articles on cancer genetics, where many articles fail to report geographical locations at all. Our pipeline combines freely available tools with rule-based processing to extract and represent relevant locations, and aims to minimize domain-customization across our two corpora. We focus on extracting locations from targeted portions of the article, including the title and any methods or study site sections deemed likely to contain relevant locations. We aim to ignore irrelevant locations, such as locations representing where certain scientific products were obtained or manufactured, not only by targeting certain text portions but also through rule-based post-processing of candidate locations. We obtain good performance, with full pipeline precision of 0.84 for the ecological and 0.78 for the biomedical corpus, allowing us to map and discuss the spatial properties of the collections.

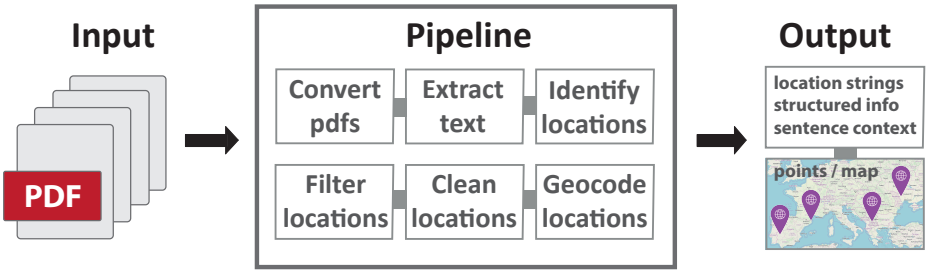


Fig 1. Overview of the processing pipeline. The pipeline starts from scientific article PDFs and outputs extracted locations, with textual and spatial representations.

Background

Extracting geographical information from scientific articles

Automatically identifying place names and their associated spatial language in text is a well-studied problem known most commonly as toponym recognition [36], and is typically the first of several steps required to map or spatially index a corpus [10]. Approaches to toponym recognition (or more broadly, location identification) in scientific articles have thus far mainly consisted of *rule-based* and *gazetteer-based* approaches [36]. A gazetteer-based approach consists of looking up words or sequences of words in a place name database (a gazetteer), where a match indicates a (likely) location. The main downside of this approach is that many common words appear in gazetteers as locations, such as 'bath', 'nice', and 'of', and hence false positives must be limited via post-processing or careful targeting of words to look up. In one example of this approach [23], sentences are first tagged with part-of-speech (POS) labels (such as 'noun', 'adjective', or 'noun phrase'), and any noun phrases containing capitalized words

are looked up in the GeoNames gazetteer and in Google Maps. A rule-based approach is used in [24] which detects patterns of relevant words, including words found in a gazetteer (likely to be a location), location modifiers (e.g 'north'), and entity type words (e.g. 'river' or 'mountain'). Good performance is obtained after adding custom pre- and post-processing steps, such as enhancing word lists with geology-specific terms and detecting citations in order to skip them as location candidates.

A commonly used method to identify toponyms in text is to run a Named Entity Recognition (NER) tool over the text and retain the subset of entities which are tagged as locations. However, out-of-the-box NER tools have often been trained mostly on news articles and their performance tends to decrease when texts diverge in form and content from these [37]. An NER tool is considered in [24] for the task of extracting geographical/geological locations in geology articles, but rejected in favor of a rule-based approach due to poor performance. In a series of papers on the aforementioned phylogeography corpus, custom NER tools are developed to identify toponyms, including first using a rule-based approach [27], followed by higher-performing machine learning models using first Conditional Random Fields (CRFs) [29], then bi-directional recurrent neural networks (RNNs) [30]. However, their custom NER tools require re-training on an annotated corpus, as opposed to out-of-the-box tools which can be more readily applied to varied corpora, and no filtering is done to identify only a relevant subset of toponyms/locations.

In this work, we use a pre-trained, freely available NER tool and combine it with rules to deliver as output a subset of relevant locations for each scientific article, such as study sites or patient treatment locations. We focus on extracting these relevant locations by targeted specific portions of the article (pre-NER processing) and by filtering candidate locations to exclude company locations and other irrelevant locations (post-NER processing).

Geographically representing scientific articles

Once locations have been identified and extracted from an article, a subsequent step is required to convert these textual locations to an explicitly spatial representation. This step is referred to as toponym resolution [38], grounding, or geocoding, and involves both resolving ambiguity (such as, determining whether the string 'Zürich' refers to the city of Zürich, the canton of Zürich, or perhaps even Zürich airport) and assigning a geometry to represent the location (such as a latitude, longitude point for the city of Zürich, Switzerland). Geometries are usually obtained by linking the extracted location to a particular gazetteer record which also contains a geometry. In practice, this step can simply consist of querying a *geocoding* service with the location string to get back a ranked list of results, including structured information and a geometry for each, typically a point representation. To aid disambiguation, additional geographical context can be given to most geocoding services, such as a bounding box or country of interest to limit the results, or an augmented string with a containing region such as a state or country. Examples of geocoding services include the Google Geocoding API[2], OpenStreetMap (OSM) Nominatim[3], and the GeoNames search webservice[4].

In previous work dealing with geographic locations in scientific articles, the toponym resolution or geocoding step is sometimes absent, with the focus still largely on developing better methods to identify the locations of interest in text [24, 29]. Furthermore, many of the works which map study sites have annotated their article collections manually and hence do not perform automatic geocoding [12, 14, 21, 22]. Of the works which perform geocoding automatically, [23] use the relevance-sorted results

[2]https://developers.google.com/maps/documentation/geocoding/intro
[3]https://wiki.openstreetmap.org/wiki/Nominatim
[4]http://www.geonames.org/export/geonames-search.html

from both GeoNames and Google Maps and look for a containing country in the same sentence as the location string, while [13] also use the Google Maps API and rely on semi-automatic post-geocoding filtering to limit the number of false matches. In [27], GeoNames search results are disambiguated using a population heuristic (choosing the result with the highest population), a distance heuristic (choosing the result which minimizes the total geographical distance to all other toponyms in the document), and a 'metadata' heuristic tailored to their phylogeographic data. In [9], location strings are linked to gazetteer records but no disambiguation is performed, which leads to many false positive matches.

In this work, we use the Google Geocoding API, a high-performing tool, to get structured information and a spatial representation for our extracted location strings. The returned information includes a fully qualified location string, a return type with granularity indications, as well as a latitude and longitude which can be mapped. We programmatically generate maps from these results for each corpus, which gives a visual overview of the overall spatial coverage of the articles.

Materials and methods

Corpora

We benefited from the use of two article corpora to work with, which had already been identified as of interest for domain-specific meta-studies:

- **Orchards**: This corpus consists of articles relating to fruit orchards, collected to conduct a meta-analysis on the impact of agricultural practices on biodiversity [39], with an intended focus on orchards in a Mediterranean climate. We obtained an early, minimally-triaged collection of articles to develop our methods. The articles are from a varied list of ecology-related journals, with the top 4 most frequent journals being 'Environmental Toxicology and Chemistry', 'Agroforestry Systems', 'Archives of Environmental Contamination and Toxicology', and 'Apidologie'.

- **Cancer**: This corpus consists of articles used in the curated cancer genomics database Progenetix[5], specifically focused on Comparative Genomic Hybridization experiments, alongside Whole Genome/Exome Sequencing studies [40]. As part of data curation, locations are manually extracted for each article, which is currently done by taking the location of the first author, rather than by manually looking through the article contents for locations such as where patient material was obtained. The top 4 most frequent journals for articles in this cancer-genetics-focused collection are: 'Genes, Chromosomes & Cancer', 'Cancer Genetics and Cytogenetics', 'Journal of Pathology', and 'Oncogene'.

We manually annotated 150 articles in total for the Orchards corpus and 200 for the Cancer corpus (Table 1). The articles were randomly chosen for annotation from a wider set of articles which, for the Cancer corpus, were in the Progenetix database and had a full PDF available, and for the Orchards corpus, had been obtained from targeted keyword searches (as described in [39]) but were not extensively triaged. For each corpora, we set aside 50 randomly sampled articles to use as a test set; our training set consisted of the remaining annotated articles, which we used to develop our processing pipeline, including methods section detection, location extraction, and location geocoding.

In addition to annotating the ground truth locations which we found in the article contents, we also systematically annotated the quality of the textual location

[5] https://progenetix.org/

Table 1. Summary information about the two corpora used.

Corpus		Articles (annotated)		
name	domain	total	train	test
Orchards	ecology	150	100	50
Cancer	biomedical	200	150	50

information and, to help develop our methods, where this information was present in the article. Our annotations show that the location reporting quality is varied in the Cancer corpus[6], but nearly always of high quality in the Orchards corpus (Fig 2 (a)). In terms of the year of publication of the articles in our two corpora, it is the Orchards corpus that shows greater variation, with articles spanning the range 1975-2016 (Fig 2 (b)); the oldest article in the Cancer corpus by comparison is from 1995, which makes sense considering the corpus' focus on particular scientific techniques which were only developed in the 1990s.

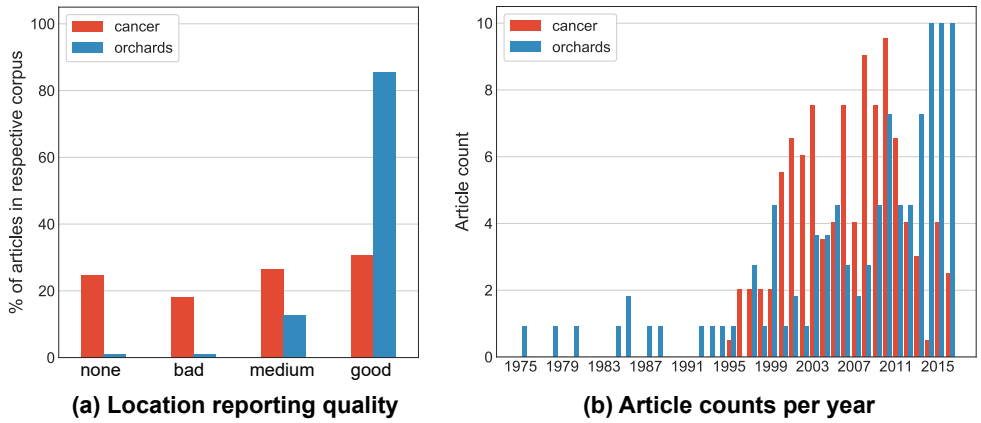


Fig 2. Comparison of the Orchards and Cancer corpora. (a) location reporting quality in the article contents, (b) publication year of the articles. Categories for location reporting quality (a): none: no mention of study/sample location; bad: implicit location info or reference to another paper; medium: study/sample location info like name of institute only and perhaps some locations not mentioned; good: explicit study/sample location info that could probably be extracted and geocoded.

Processing pipeline

We now describe the steps of the processing pipeline we applied to the two corpora, followed by any corpus-specific customizations we made to our code[7]. In general, we tried to limit extracted locations to *relevant* locations in two ways: 1. by only looking for locations in targeted portions of the article (pre-NER *Extract text* step) and 2. by filtering identified locations (post-NER *Filter locations* step).

- **Convert PDFs:** The pipeline starts from a set of PDF documents, and converts each document to 1. a plain text file, using pdfminer[8], and to 2. an XML file using CERMINE [41], a Java-based library to extract metadata and contents from scientific article PDFs. Performing two independent file conversions means the

[6]We report further on the location reporting quality over time in S1 Fig.
[7]Our code alongside article information is available at https://github.com/eacheson/pyscine
[8]https://github.com/pdfminer/pdfminer.six

pipeline has the possibility to recover from a failed XML conversion or from insufficient headings in the XML file.

- **Extract text:** The next step targets portions of the article contents in which to look for location information; this is done by identifying relevant headings (such as methods or study site sections) using regular expression matching. Matches are found by testing each paragraph beginning in the text files and each heading in the XML files. When a match is found, paragraphs under the matched heading are stored for the next step. At the end of this step, the pipeline continues using only the XML files, unless relevant headings/text were identified in the text files and not in the XML file. The article title identified in the XML file is also separately retained for the further step.

- **Identify locations:** The text portions extracted in the previous step are now processed for locations. First the text is split into paragraphs, normalized (accented characters), split into sentences and words, and a part-of-speech (POS) tagger is run over each sentence. The text is now ready for NER, which is performed using Stanford NER [42], accessed from the NLTK python library [43] (Stanford NER v3.8.0, NLTK v3.2.5). A 3-class classifier is used which tags each word as one of 'location', 'person', 'organization', or 'other' meaning not a named entity. These token (word, tag) combinations from the NER output are processed using custom code which retains sequences of tokens as location candidates which will then be triaged. The goal of this step is high recall, that is, to miss as few true location descriptions as possible. Accordingly, we keep any sequence of words with at least one named entity and include within these sequences words that often appear within a location string, such as 'in' or 'upon' and two-letter state abbreviations.

- **Filter locations:** The location candidates identified in the previous step are now filtered using rules to remove any candidate that is not deemed a relevant location, including non-locations, suspected company locations, and citations. The rules in this step were developed iteratively on the training set and are based on: tag sequences (e.g. reject candidates with no 'location' tags), presence of keep words (e.g. keep candidates with 'University' or 'Institute'), presence of discard words (e.g. reject candidates with 'Inc' or 'GmbH'), and token (tag, word) combinations. The goal of this step is to increase precision, while trying to maintain good recall. This step produces our final list of identified content locations.

- **Clean locations:** Each content location string retained in the previous step is cleaned of any trailing prepositions or punctuation before the geocoding step.

- **Geocode locations:** Each clean location string is sent to the Google Geocoding API, and the top result is retained (if any results are returned). Each geocode result provides structured location information, including a qualified string representation of the location (such as 'San Francisco, CA, USA' for the query 'San Francisco'), a latitude, and a longitude.

The processing pipeline is illustrated in Fig 3. Note that whenever a location candidate is retained, the sentence it was found in is also retained, so that the final output consists not only of identified content location strings and their geocode result information, but also of their sentence context. This not only facilitates our own evaluation, but allows for complex compositional location descriptions (such as '30 km from Florence, Italy') and coordinates appearing in text (such as 'Florence, Italy (43.77° N, 11.26°E)') to be retained in our structured output for a human annotator to easily access, as these are typically in the same sentence as a location that our pipeline does

retain (such as 'Florence, Italy' in both previous examples). Note that we adapted and ran the coordinate parsing code from [26], but it performed poorly on our data because coordinate strings were often transformed erroneously during conversion from PDFs to plain text and XML files.

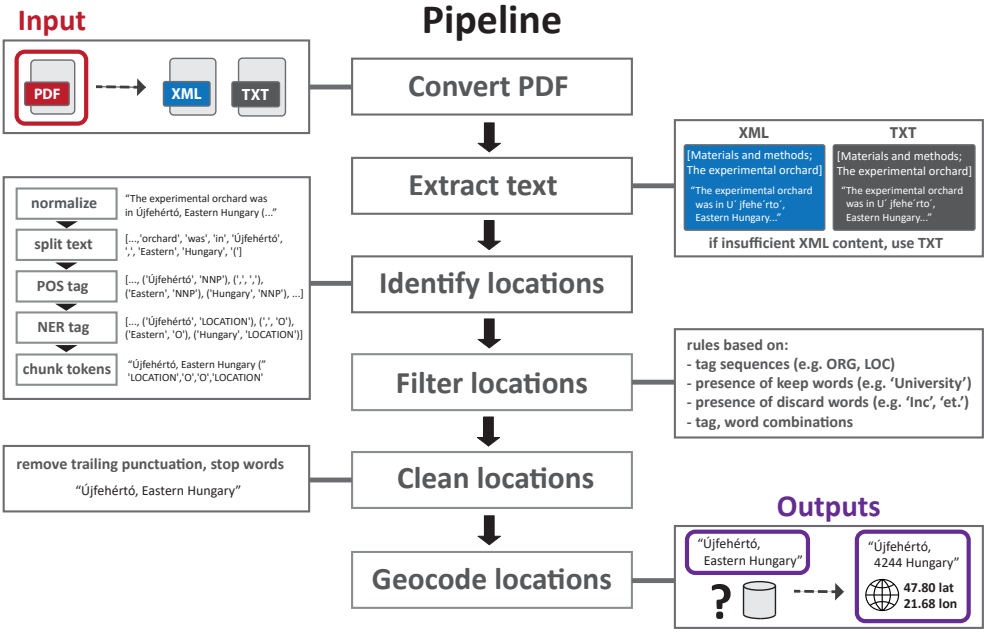


Fig 3. Detailed view of the article processing pipeline. Pipeline illustrated using an example from the Orchards corpus.

We minimally customized our pipeline for either the Orchards or the Cancer corpus. The first and most important customization was in the regular expressions used to detect relevant section headings (*Extract text* step). Relevant headings in the Orchards corpus featured words like 'region', 'area', and 'site', whereas in the Cancer corpus, words indicative of a relevant section heading included 'patient', 'sample', 'specimen', and 'subject'. The second customization was in the rules used to retain certain sequences of tokens as location candidates (*Identify locations* step). In the Orchards corpus, location strings often contained cardinal direction words (such as 'east', 'southern', or 'northeastern') as well as geographic entity type words (like 'region', 'county', and 'park'). We found that including these words in our final location strings had an overall positive effect on the geocoding step, mainly because it tended to keep location words describing the same location together as one string as opposed to two distinct strings (such as 'Nancy (East of France)' instead of 'Nancy' and 'France'), giving better context for the geocoding step.

Results

Our pipeline produced two main outputs: extracted location strings and geocode results. In addition, evaluation could be performed against two slightly different units: location units or article units. In order to evaluate our two main outputs separately as well as in sequence, we first evaluated our pipeline in 3 stages, using the location unit (Table 2): 1. first, we calculated whether each extracted string was correct (a true positive) or not, giving a value for extraction precision; 2. we then separately evaluated the geocoding

using the subset of true positive extracted location strings by calculating how often the geocode result for these strings was correct or incorrect, giving a value for geocoding accuracy; 3. we finally looked at the full set of extracted location strings (true and false positives) and evaluated the final geocode result for each, giving a value for full pipeline precision. This full pipeline evaluation includes several cases where the final result is worse than the individual steps (a correct location string was extracted, but geocoded to the wrong location, a false positive overall), but also a few cases where the full pipeline is better than the individual steps (a wrongly extracted location had no geocode result, resulting in a true negative overall). This is reflected in Table 2, where the full pipeline precision is slightly lower than the extraction precision for both corpora[9].

In a second evaluation, we evaluated extraction precision, recall, and $F1$ using the article unit, in order to not give a disproportionate amount of weight to articles with multiple study sites or sample locations. Specifically, we calculated both precision and recall out of a maximum value of 1 for each article, where a precision of 1 meant all extracted location strings were correct, and a recall of 1 meant all ground truth locations (e.g. study sites) were represented in the extracted strings. We then summed these values for an overall precision and overall recall, respectively (Table 2). Precision and recall were combined into one value, $F1$, through their harmonic mean. Any locations extracted from the title were included in this overall pipeline evaluation for the Orchards corpus, as it was determined at the training stage that the titles in this corpus, but not in the Cancer corpus, contained useful locations.

Table 2. Results for both corpora, organized according to whether the location unit or article unit was used in evaluation.

	location unit			article unit		
	extraction	geocoding	full pipeline	extraction (weighted)		
corpus	precision	accuracy	precision	precision	recall	F1
Orchards	0.869	0.906	0.842	0.822	0.804	0.813
Cancer	0.810	0.980	0.778	0.740	0.769	0.754

Indeed, in the Orchards test corpus, 19 titles contained a location in the title, whereas in the Cancer test corpus, just one title arguably contained a location, but in adjectival form (e.g. 'Korean tumours'). We achieved very good performance on title extraction in the Orchards test corpus, with 0.95 for both precision (18/19) and recall (18/19), and hence also $F1$. These locations were often a good overall summary of the study region, but were also fairly often vague regions: in this test set, 6 out of 19 ground truth locations had some inherent vagueness (examples include 'Southern Russia', 'eastern Spain', 'European Alps').

Our results in Table 2 show that generally performance was superior on the Orchards corpus, which is consistent with the superior location reporting quality in that corpus (Fig 2 (a)). However, the geocoding accuracy was higher for the Cancer corpus. Though both corpora often had location strings which weren't fully qualified with a city or country, the Google Geocoding API still mostly gave correct answers for unqualified strings in the Cancer corpus (e.g. 'Massachusetts General Hospital', 'Royal Free Hospital and Medical School') but not in the Orchards corpus (e.g. 'Via Emilia', 'Dry Creek Vineyard'). Indeed, generally the Orchards corpus featured study sites in lesser known locations outside of cities, whereas the Cancer corpus featured more well-known location names such as cities in Europe and North America, and large hospitals or research Universities.

[9]For the Orchards corpus, we present the results on a subset of articles consisting of studies, rather than review articles, editorials, or articles in popular science magazines. These studies formed between 73-74% of articles in the full minimally-triaged collection, training set, and test set. For results on the complete Orchards corpus, see the S1 Table in supplementary materials.

We systematically classified the errors in our pipeline based on the 3-stage evaluation results (Table 3 and Table 4). Only the main source of error for each location unit was recorded and only when the full pipeline result was incorrect did we record an error. NER errors were the most frequent kind of error, followed by not having extracted the paragraph or sentence containing the location string (hence not making it to the NER step). The 'comma group' errors occurred when there were multiple, separate locations separated by commas, which our code chunked together as a single qualified location (e.g. 'Burlington, Cambridge' where Burlington and Cambridge were separate towns in Canada, instead of Burlington being contained by Cambridge). Thankfully these comma group errors were all in the Orchards corpus and 3 of them were in one article which listed several countries one after another, something which could be adjusted in code by detecting comma-separated countries.

Table 3. Errors in both corpora classified into categories. Errors shown as raw counts and as the percentage of the total errors for that corpus.

	Orchards		Cancer	
error description	count	percent	count	percent
NER error	12	27.3	8	32.0
text portion not extracted	8	18.2	7	28.0
wrong/no geocode result	9	20.5	1	4.0
comma group	7	15.9	0	0.0
candidate filtering error	3	6.8	4	16.0
non-standard headings	3	6.8	0	0.0
other	2	4.5	5	20.0
total	44	100	25	100

Table 4. Examples for each error category.

error description	example
NER error	Rome tagged as location in 'MacIntosh or Rome varieties'
text portion not extracted	location only appears in Acknowledgements
wrong/no geocode result	'Moldova Region' in Romania geocoded to Moldova country
comma group	'Burlington, Cambridge' taken as one location
candidate filtering error	company location not filtered out
non-standard headings	'Almonds' sub-heading contained study site info
other	wrongly extracted publisher location in footer

Fig 4 illustrates the spatial distribution of geocoded locations extracted from our two corpora at a global scale. Any extracted string which gave a geocode result is mapped, and hence the color-coding represents the full pipeline precision (c.f. full pipeline precision column found in Table 2). Note that, especially in the Cancer corpus, the majority of full pipeline false positives are due to wrongly extracted locations (extraction false positive), rather than geocoding errors. Hence the same map without color-coding would represent what one would see when mapping a new, unevaluated corpus. Both maps are dominated by locations in Europe and North America, demonstrating underlying geographic properties of these corpora. For the Orchards corpus, locations around the Mediterranean reflect the underlying intent of the corpus. In the Cancer corpus, the locations identified suggest facilities capable of carrying out sophisticated genetic analysis of cancers. In both maps, false positives are predominately found in North America, likely reflecting both biases in the underlying spatial data used in geocoding and an underlying tendency of the geocoder to default to locations in North America.

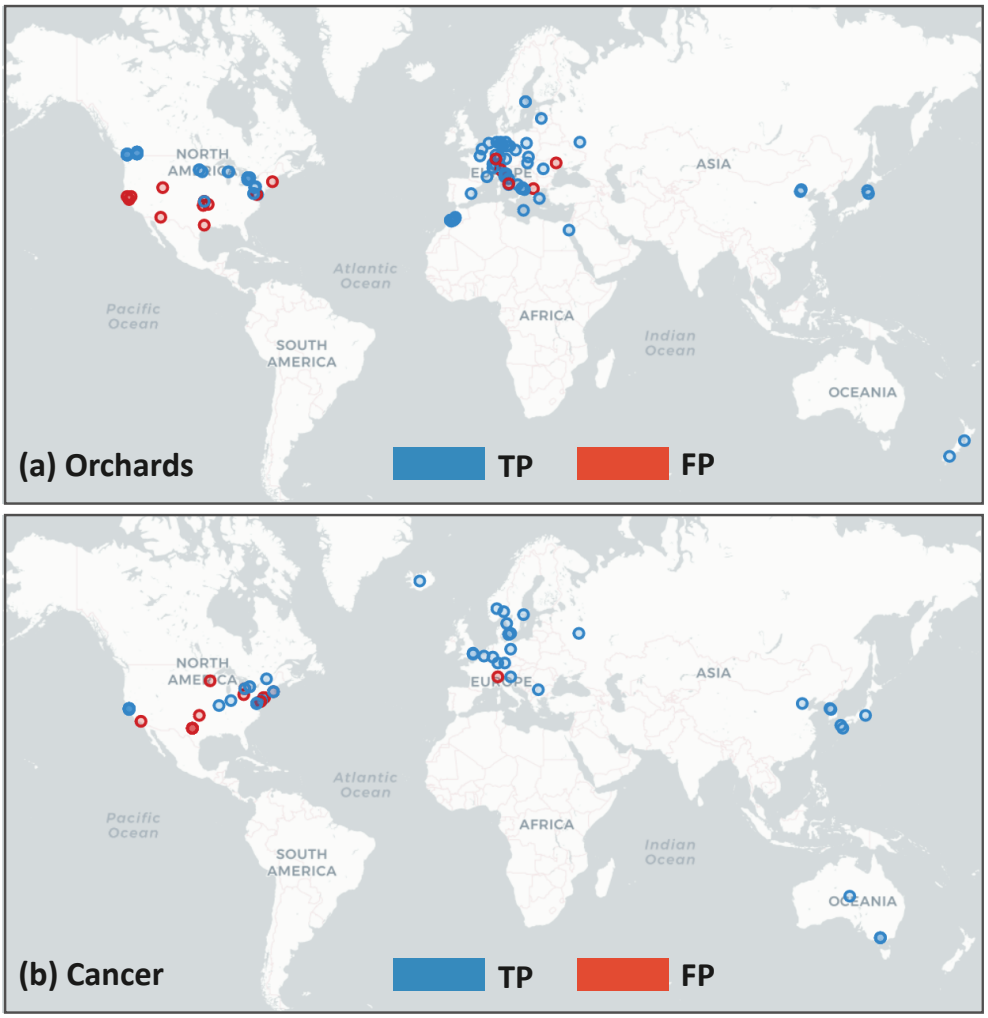


Fig 4. Global maps of geocoded locations. (a) Orchards corpus and (b) Cancer corpus. In both maps, full pipeline precision is represented (that is, any extracted string which also gave a geocode result is mapped), with true positives (TP) in blue and false positives (FP) in red.

Discussion

In this work, we sought to automatically extract and represent meaningful locations from scientific articles from both the ecology and biomedical domains. Relatively few works have been published on this specific problem and, of the works that share such an aim, the majority have focused on the ecological domain [12–14, 21, 23, 26], with two works examining a slightly broader set of journals still focused on environmental research [22, 26], one studying geology articles [24], and one in the hydrology/ hydrogeology domain [44]. In many of these works, location identification/extraction from text is performed manually [12, 14, 21, 22] or semi-automatically [13]. Our work shows that it is possible to build a fully automated pipeline, with limited customization across research domains within the broader text type of scientific articles, and obtain results of a high enough quality to be useful in the context of a meta-analysis or of a geographical search/filter for articles.

An important limitation of some current work is a tendency to develop customized tools for particular tasks and corpora [29,30,45]. We deliberately set out to build a more generic pipeline, whose focus lay on identifying relevant locations from scientific articles using existing tools. Our approach therefore does not aim to optimize individual components of the pipeline (e.g. NER for toponym recognition or geocoding for toponym resolution), but rather aims to provide a useful set of filtered locations which can then be subject to human analysis. To facilitate this, our pipeline delivers locations in multiple formats (location strings, point coordinates, and location sentences), suitable for review and correction by a human annotator to further increase the overall precision, particularly using the location sentences. Confidence or uncertainty scores could also be assigned to each article, such as is done in [46] where a baseline score is increased or decreased based on the intermediate outputs of a rule-based pipeline. Finally, our task and pipeline leads to output that is more manageable for a human annotator (e.g. in the context of a geographical corpus analysis), because we focus precisely on those locations that would be the main content locations for an article.

Although we aimed to develop a generic pipeline, we did include some elements of customization. In particular for the Orchards corpus, we attempt to extract more than location names by, for example, including cardinal direction terms. However, we make no attempt to extract truly compositional place descriptions such as '30km from Florence' or interpreting these descriptions, though our code could be adapted to recognize these types of expressions and could be given to a system similar to the one in [46] used to georeference location descriptions for animal specimens. However, even if such expressions were extracted with high precision, current geocoding tools typically do not handle such expressions, despite long-standing calls to do so [36].

One important limitation of our work is the representation of all extracted locations as points. Although this is justified in most cases at a global scale, this may quickly become inappropriate depending on the properties of a particular corpus. Depending on our viewpoint and purpose, the Cancer corpus could be used to analyse locations related to the genetic analysis of tumour data (where point representations, related to specific facilities, are appropriate) or to explore locations related to tumour incidence (where more aggregated locations, related to large regions served by specialized hospitals, would be more meaningful). Indeed we are largely constrained to the use of points to initially represent all extracted locations, given points are returned by the geocoding service, but we could also use a bounding box for a subset of results, which gives an indication of area. Importantly, by keeping location representations in both textual and explicitly spatial form, there remains the possibility of re-generating and refining geometries using the extracted location strings.

Though a point is a rather simplistic way to represent a single scientific article, a larger collection of such points may be a good way to represent and map an entire corpus of articles, particularly on the global scale where small differences in study site areas would not be visible. Global density maps, such as the kernel density map of sites in [14] or the rectangular-grid point aggregation in [26], can be created from point collections and are especially useful to highlight geographical research gaps in the corpus as a whole. As for interactive maps of study sites, a good example is JournalMap[10], a geosemantic search tool developed for an ecology-focused corpus where locations have been manually identified [21]. One straightforward enhancement of this tool would be to use bounding boxes to estimate the area/scale of study sites.

An alternative approach to performing our task would be to use sentence classification, including recently developed deep learning approaches [47]. Instead of identifying a set of relevant locations by targeting certain portions of the article (pre-NER) and filtering irrelevant locations (post-NER), one could instead classify each

[10] https://www.journalmap.org/

sentence in the article as either describing study/sample sites or not. Those sentences likely to contain study/sample site descriptions could then be further processed to extract location strings to be sent to a geocoding tool, such as is done in our work. Such a classification approach was used in [23] who classify sentences into 'environmental' or 'experimental' sentences, with the environmental sentences featuring relevant locations such as study sites, and experimental ones featuring irrelevant locations such as the provenance of chemicals.

Conclusion

Writing a scientific paper is time-consuming and expensive, and we should maximize the value of each and every scientific work. Full-text analysis on large article collections is now possible, and should be increasingly applicable thanks to open access policies making more full-text articles available for processing. In this paper, we processed two collections of scientific articles, starting from collections of full-text PDFs, extracting locations using NER tools and rule-based processing, and geocoding these locations to spatially represent them.

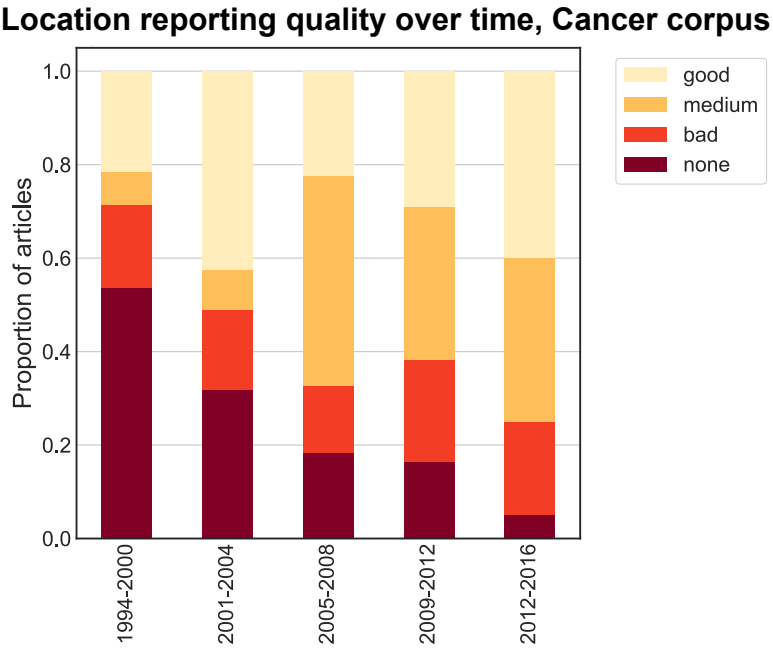
Recording spatially explicit geographical information (such as a point coordinate, a bounding box, or a set of geometries) for scientific articles is an important step to facilitate meta-analyses and to identify geographical biases in scientific research. We tackled this problem by building an automatic processing pipeline, with the following takeaways:

- We use current tools, with minimal customization, making our pipeline easily extendable to other corpora of scientific articles.
- Our pipeline has high precision for identifying and resolving relevant location mentions (0.84 for an environmental corpus and 0.78 for a biomedical one) and is effective in extracting relevant locations at the article level (F1 0.81 for the environmental corpus and 0.75 for the biomedical one).
- We specify our task such that the aim is to filter and identify only relevant location mentions, suitable for both visualization and processing by human annotators. We reduce the number of location mentions to be triaged greatly through our approach.
- An error analysis reveals that failures can occur throughout the chain. These failures are also dependent on the nature of the problem specification (e.g. the difference between identifying all toponyms or identifying relevant location mentions).

Future systems will benefit from improvements in the performance of individual system components (e.g. improved toponym recognition through deep learning approaches). Equally, the ability of geocoders to return more complex geometries, as appropriate for the scale of analysis, has clear potential for both representation and analysis of scientific corpora. We suggest that future work focus not only on such improvements in individual tasks, but also on gathering requirements from potential users of geographical exploration and search interfaces for scientific article corpora. The success of these approaches depends on their usefulness and practicality.

Supporting information

S1 Fig. Location reporting quality over time for the Cancer corpus. We combined our location quality judgements with the manually annotated publishing years of all our manually annotated articles ($N = 199$, one article was excluded because it contained no samples and instead developed an algorithm) to plot the evolution of location quality reporting over time. For each time interval, we plotted the proportion of articles in that time interval which were in each of 4 location quality categories (good, medium, bad, none). The resulting plot suggests that location reporting quality is slowly improving over time. In particular, the proportion of articles reporting no location at all is steadily decreasing and the proportion of articles with either 'good' or 'medium' location reporting is trending upwards.



S1 Table. Full results for the Orchards corpus. Below, we show both the results considering just studies ('Orchards-studies') and the results for the full set of articles ('Orchards-full'), including other article types (e.g. reviews, editorials, and popular science articles).

	location unit			article unit		
	extraction	geocoding	full pipeline	extraction (weighted)		
corpus	precision	accuracy	precision	precision	recall	F1
Orchards-studies	0.869	0.906	0.842	0.822	0.804	0.813
Orchards-full	0.872	0.908	0.846	0.836	0.760	0.797

Acknowledgments

Many thanks to Markus van der Meer for direct access to PDFs forming the Orchards corpus and to Paula Carrio Cordo and Michael Baudis for direct access to PDFs forming the Cancer corpus.

References

1. Teitler BE, Lieberman MD, Panozzo D, Sankaranarayanan J, Samet H, Sperling J. NewsStand: A New View on News. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. GIS '08. New York, NY, USA: ACM; 2008. p. 18:1–18:10. Available from: http://doi.acm.org/10.1145/1463434.1463458.

2. Buscaldi D, Magnini B. Grounding Toponyms in an Italian Local News Corpus. In: Proceedings of the 6th Workshop on Geographic Information Retrieval. GIR '10. New York, NY, USA: ACM; 2010. p. 15:1–15:5. Available from: http://doi.acm.org/10.1145/1722080.1722099.

3. D'Ignazio C, Bhargava R, Zuckerman E, Beck L. Cliff-clavin: Determining geographic focus for news. NewsKDD: Data Science for News Publishing, at KDD 2014. 2014;.

4. Dredze M, Paul MJ, Bergsma S, Tran H. Carmen: A twitter geolocation system with applications to public health. In: AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI). Citeseer; 2013. p. 20–24. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.309.6126&rep=rep1&type=pdf.

5. Zhang W, Gelernter J. Geocoding location expressions in Twitter messages: A preference learning method. Journal of Spatial Information Science. 2014;(9). doi:10.5311/JOSIS.2014.9.170.

6. Middleton S, Kordopatis-Zilos G, Papadopoulos S, Kompatsiaris I. Location extraction from Social Media: geoparsing, location disambiguation and geotagging. ACM Transactions on Information Systems. 2018;.

7. Shapiro JT, Báldi A. Lost locations and the (ir)repeatability of ecological studies. Frontiers in Ecology and the Environment. 2012;10(5):235–236. doi:10.1890/12.WB.015.

8. Karl JW, Gillan JK, Herrick JE. Geographic searching for ecological studies: a new frontier. Trends in Ecology & Evolution. 2013;28(7):383–384. doi:10.1016/j.tree.2013.05.001.

9. Kmoch A, Uuemaa E, Klug H, Cameron SG. Enhancing Location-Related Hydrogeological Knowledge. ISPRS International Journal of Geo-Information. 2018;7(4):132. doi:10.3390/ijgi7040132.

10. Jones CB, Purves RS. Geographical information retrieval. International Journal of Geographical Information Science. 2008;22(3):219–228. doi:10.1080/13658810701626343.

11. Gerstner K, Moreno-Mateos D, Gurevitch J, Beckmann M, Kambach S, Jones HP, et al. Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. Methods in Ecology and Evolution. 2017;8(6):777–784. doi:10.1111/2041-210X.12758.

12. Wallis PJ, Nally RM, Langford J. Mapping Local-Scale Ecological Research to Aid Management at Landscape Scales: Mapping Ecological Research at Landscape Scales. Geographical Research. 2011;49(2):203–216. doi:10.1111/j.1745-5871.2011.00691.x.

13. Fisher R, Radford BT, Knowlton N, Brainard RE, Michaelis FB, Caley MJ. Global mismatch between research effort and conservation needs of tropical coral reefs. Conservation Letters. 2011;4(1):64–72. doi:10.1111/j.1755-263X.2010.00146.x.

14. Martin LJ, Blossey B, Ellis E. Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. Frontiers in Ecology and the Environment. 2012;10(4):195–201. doi:10.1890/110154.

15. Frenken K, Hardeman S, Hoekman J. Spatial scientometrics: Towards a cumulative research program. Journal of Informetrics. 2009;3(3):222–232. doi:10.1016/j.joi.2009.03.005.

16. Pan RK, Kaski K, Fortunato S. World citation and collaboration networks: uncovering the role of geography in science. Scientific Reports. 2012;2:902. doi:10.1038/srep00902.

17. Fried D, Kobourov SG. Maps of Computer Science. In: 2014 IEEE Pacific Visualization Symposium. Yokohama: IEEE; 2014. p. 113–120. Available from: `http://ieeexplore.ieee.org/document/6787157/`.

18. Korhonen A, Ó Séaghdha D, Silins I, Sun L, Högberg J, Stenius U. Text Mining for Literature Review and Knowledge Discovery in Cancer Risk Assessment and Research. PLOS ONE. 2012;7(4):e33427. doi:10.1371/journal.pone.0033427.

19. Simpson MS, Demner-Fushman D. Biomedical Text Mining: A Survey of Recent Progress. In: Aggarwal CC, Zhai C, editors. Mining Text Data. Boston, MA: Springer US; 2012. p. 465–517. Available from: `https://doi.org/10.1007/978-1-4614-3223-4_14`.

20. Frenken K, Hoekman J. Spatial Scientometrics and Scholarly Impact: A Review of Recent Studies, Tools, and Methods. In: Ding Y, Rousseau R, Wolfram D, editors. Measuring Scholarly Impact: Methods and Practice. Cham: Springer International Publishing; 2014. p. 127–146. Available from: `https://doi.org/10.1007/978-3-319-10377-8_6`.

21. Karl JW, Herrick JE, Unnasch RS, Gillan JK, Ellis EC, Lutters WG, et al. Discovering Ecologically Relevant Knowledge from Published Studies through Geosemantic Searching. BioScience. 2013;63(8):674–682. doi:10.1525/bio.2013.63.8.10.

22. Margulies JD, Magliocca NR, Schmill MD, Ellis EC. Ambiguous Geographies: Connecting Case Study Knowledge with Global Change Science. Annals of the American Association of Geographers. 2016;106(3):572–596. doi:10.1080/24694452.2016.1142857.

23. Tamames J, de Lorenzo V. EnvMine: A text-mining system for the automatic extraction of contextual information. BMC Bioinformatics. 2010;11:294. doi:10.1186/1471-2105-11-294.

24. Leveling J. Tagging of Temporal Expressions and Geological Features in Scientific Articles. In: Proceedings of the 9th Workshop on Geographic Information Retrieval. GIR '15. New York, NY, USA: ACM; 2015. p. 6:1–6:10. Available from: `http://doi.acm.org/10.1145/2837689.2837701`.

25. Page RDM. Enhanced display of scientific articles using extended metadata. Web Semantics: Science, Services and Agents on the World Wide Web. 2010;8(2):190–195. doi:10.1016/j.websem.2010.03.004.

26. Karl JW. Mining location information from life- and earth-sciences studies to facilitate knowledge discovery. Journal of Librarianship and Information Science. 2018; p. 0961000618759413. doi:10.1177/0961000618759413.

27. Weissenbacher D, Tahsin T, Beard R, Figaro M, Rivera R, Scotch M, et al. Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. Bioinformatics. 2015;31(12):i348–i356. doi:10.1093/bioinformatics/btv259.

28. Tahsin T, Weissenbacher D, Rivera R, Beard R, Firago M, Wallstrom G, et al. A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records. Journal of the American Medical Informatics Association. 2016;23(5):934–941. doi:10.1093/jamia/ocv172.

29. Weissenbacher D, Sarker A, Tahsin T, Scotch M, Gonzalez G. Extracting geographic locations from the literature for virus phylogeography using supervised and distant supervision methods. AMIA Summits on Translational Science Proceedings. 2017;2017:114–122.

30. Magge A, Weissenbacher D, Sarker A, Scotch M, Gonzalez-Hernandez G. Bi-directional Recurrent Neural Network Models for Geographic Location Extraction in Biomedical Literature. In: Biocomputing 2019. WORLD SCIENTIFIC; 2018. p. 100–111. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789813279827_0010.

31. Weissenbacher D, Magge A, O'Connor K, Scotch M, Gonzalez G. SemEval-2019 Task 12: Toponym Resolution in Scientific Papers. In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Minneapolis, Minnesota, USA; 2019. p. 907–916. Available from: https://www.aclweb.org/anthology/papers/S/S19/S19-2155/.

32. Amitay E, Har'El N, Sivan R, Soffer A. Web-a-where: Geotagging Web Content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '04. New York, NY, USA: ACM; 2004. p. 273–280. Available from: http://doi.acm.org/10.1145/1008992.1009040.

33. Anastacio I, Martins B, Calado P. A Comparison of Different Approaches for Assigning Geographic Scopes to Documents. 2009;.

34. Monteiro BR, Davis Jr CA, Fonseca F. A survey on the geographic scope of textual documents. Computers & Geosciences. 2016;doi:10.1016/j.cageo.2016.07.017.

35. Leidner JL. Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names. Edinburgh University. Edinburgh, Scotlan; 2007. Available from: http://dl.acm.org/citation.cfm?id=1328989.

36. Leidner JL, Lieberman MD. Detecting geographical references in the form of place names and associated spatial natural language. SIGSPATIAL Special. 2011;3(2):5–11.

37. Augenstein I, Derczynski L, Bontcheva K. Generalisation in named entity recognition: A quantitative analysis. Computer Speech & Language. 2017;44:61–83. doi:10.1016/j.csl.2017.01.012.

38. Leidner JL, others. Toponym resolution in text:"Which Sheffield is it?". In: Proceedings of the the 27th annual international ACM SIGIR conference (SIGIR 2004). Citeseer; 2004. p. 602. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.6004&rep=rep1&type=pdf.

39. van der Meer M, Lüscher G, Kay S, Jeanneret P. What evidence exists on the impact of agricultural practices in fruit orchards on biodiversity indicator species groups? A systematic map protocol. Environmental Evidence. 2017;6:14. doi:10.1186/s13750-017-0091-1.

40. Cai H, Kumar N, Ai N, Gupta S, Rath P, Baudis M. Progenetix: 12 years of oncogenomic data curation. Nucleic Acids Research. 2014;42(Database issue):D1055–1062. doi:10.1093/nar/gkt1108.

41. Tkaczyk D, Szostek P, Fedoryszak M, Dendek PJ, Bolikowski L. CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR). 2015;18(4):317–335. doi:10.1007/s10032-015-0249-8.

42. Finkel JR, Grenager T, Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics; 2005. p. 363–370. Available from: https://doi.org/10.3115/1219840.1219885.

43. Bird S, Loper E, Klein E. Natural Language Processing with Python. O'Reilly Media; 2009. Available from: http://shop.oreilly.com/product/9780596516499.do.

44. Kmoch A, Uuemaa E. Geo-referencing of journal articles and platform design for spatial query capabilities; 2018. Available from: https://zenodo.figshare.com/articles/Geo-referencing_of_journal_articles_and_platform_design_for_spatial_query_capabilities/6893945.

45. Li H, Wang M, Baldwin T, Tomko M, Vasardani M. UniMelb at SemEval-2019 Task 12: Multi-model combination for toponym resolution. In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Minneapolis, Minnesota, USA; 2019. p. 1313–1318. Available from: https://www.aclweb.org/anthology/papers/S/S19/S19-2231/.

46. van Erp M, Hensel R, Ceolin D, Meij Mvd. Georeferencing Animal Specimen Datasets. Transactions in GIS. 2015;19(4):563–581. doi:10.1111/tgis.12110.

47. Kim Y. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1746–1751. Available from: http://aclweb.org/anthology/D14-1181.

Appendices

Curriculum Vitae

Personal details

Name: Elise Anne Acheson
Date of Birth: November 10th 1982

Education

2015 – 2019 **PhD**, **Department of Geography, University of Zurich**
Thesis: 'Extracting and modeling the geography of text documents'
Supervised by Dr. Ross Purves, Geocomputation Unit.

2010 – 2011 **MSc., School of Informatics, University of Edinburgh (UK)**
Thesis: 'Evaluating pronoun translation for statistical machine translation'
Supervised by Dr. Bonnie Webber.

2004 – 2007 **BSc., University of British Columbia (Vancouver, Canada)**
Interdisciplinary studies in Cognitive Systems: Psychology, Computer Science, Linguistics, Philosophy. Specialization in Psychology.

Work Experience

2015 – 2019 **Graduate Researcher and Teaching Assistant, University of Zurich**
Including assisting with GIS course and Java programming course.
Co-organized "Hands-on reproducible research" workshop.

2011 – 2015 **Product Engineer, ESRI (Edinburgh, UK)**
Java software development for leading GIS software company.

Publications

Acheson, E., Volpi, M., and Purves, R. S. (2019). Machine learning for cross-gazetteer matching of natural features. *International Journal of Geographical Information Science*, 1–27.

Wartmann, F. M., Acheson, E., and Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach. *International Journal of Geographical Information Science*, 32(8):1572-1592.

Brunner, M. I., Pool, S., Kiewiet, L., Acheson, E. (2018). The other's perception of a streamflow sample: from a bottle of water to a data point. *Hydrological Processes*:1-6.

Acheson, E., Villette, J., Volpi, M., Purves, R. S. (2017). Gazetteer matching for natural features in Switzerland. In: *Proceedings of the 11th Workshop on Geographic Information Retrieval*, Heidelberg (Germany), 30 November 2017 - 1 December 2017.

Acheson, E., Wartmann, F. M., Purves, R. S. (2017). Generating spatial footprints from hiking blogs. In: *COSIT 2017: International Conference on Spatial Information Theory*, L'Aquila (Italy), 4 September 2017 - 8 September 2017, 5-7.

Acheson, E., De Sabbata, S., Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309-320.

Acheson, E., and Purves, R. S. (2016). Exploring Strava segments as a source of placename information. *GeoSocial: Social Media and GIScience* workshop at GIScience 2016, Montreal (Canada), 27 September 2016.

De Sabbata, S., Acheson, E. (2016). Geographies of gazetteers in Great Britain. In: *24th GIS Research UK* (GISRUK 2016) conference, Greenwich, 30 March 2016 - 1 April 2016.